How to Select Datapoints for Efficient Human Evaluation of NLG Models?

Vilém Zouhar Peng Cui Mrinmaya Sachan
Department of Computer Science, ETH Zurich, Switzerland
{vzouhar, pcui, msachan}@ethz.ch

Abstract

Human evaluation is the gold standard for evaluating text generation models. However, it is expensive. In order to fit budgetary constraints, a random subset of the test data is often chosen in practice for human evaluation. However, randomly selected data may not accurately represent test performance, making this approach economically inefficient for model comparison. Thus, in this work, we develop and analyze a suite of selectors to get the most informative datapoints for human evaluation, taking the evaluation costs into account. We show that selectors based on variance in automated metric scores, diversity in model outputs, or Item Response Theory outperform random selection. We further develop an approach to distill these selectors to the scenario where the model outputs are not yet available. In particular, we introduce source-based estimators, which predict item usefulness for human evaluation just based on the source texts. We demonstrate the efficacy of our selectors in two common NLG tasks, machine translation and summarization, and show that only ∼70% of the test data is needed to produce the same evaluation result as the entire data.

1 Introduction

Robust model evaluation drives natural language generation (NLG) research. Despite improvements in automated evaluation metrics, human evaluation remains the gold standard in NLG (Zhou et al., 2022; Freitag et al., 2023). Distinguishing between high-quality models requires increasingly more expert human annotations to determine which model performs better. For example, WMT shared tasks (Conference on Machine Translation, Kocmi et al., 2024a), annually evaluate dozens of state-of-the-art machine translation models.

The dominant approach in human evaluation under budgetary constraints is to *randomly* select evaluation items in the test data, as shown by the survey of Ruan et al. (2024). The random selection is clearly suboptimal because it can lead to selecting overlapping items while omitting items that impact the evaluation outcome (e.g., model ranking) more.

Thus, in this paper, we develop approaches to select a subset of test items to reduce human-evaluation cost without sacrificing evaluation accuracy. In particular, we are interested in selecting a subset of test items that would lead to a human ranking of multiple NLG models similar to that on the entire test set. We frame this as a subset selection problem (dubbed output-based selection, Figure 1) where we are given a set of items $\mathcal{X}$ and model outputs $\mathcal{M}$, and would like to select a subset $\mathcal{Y} \subseteq \mathcal{X}$ such that the ranking of models with human evaluation on $\mathcal{Y}$ is the same as on $\mathcal{X}$. Although automated metrics can be noisy and may not always align with human judgments, they still help identify test items that are informative for evaluation. We build multiple data selectors that leverage these metrics, such as prioritizing challenging items, items leading to diverse model outputs, or using Item Response Theory.

However, in some evaluation scenarios, such as when organizing a large shared task, we may not yet know which models will be evaluated by our test set, or obtaining all model outputs on the whole $\mathcal{X}$ may be computationally infeasible due to the size of $\mathcal{X}$. In this setting, called source-based selection (Figure 3), the standard methods for output-based selection cannot be used. However, by distilling the output-based selection methods, we build predictors that only use the item input to predict the expected item difficulty or likelihood that it leads to diverse model outputs.

We demonstrate the efficacy of our data selection approach with case studies on two typical natural language generation tasks: machine translation and summarization. Our key contributions include:

- framing the task of informative subset selection for evaluation in two variants,

[0]We release the subset2evaluate package, trained models, and code for reproducing the results in this paper.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 1789–1811, 2025. https://doi.org/10.1162/TACL.a.60
Action Editor: André F.T. Martins. Submission batch: 2/2025; Revision batch: 5/2025; Published 12/2025.

- multiple evaluation subset selection methods, including cost- and document-aware selection,
- selector distillation for source-based selection,
- package subset2evaluate for budget-efficient test set construction for model evaluation.

This paper is structured as follows. We formalize the problem in Section 2, and describe the output-based and source-based selectors in Sections 3 and 4. We show our results on efficient machine translation evaluation in Section 5 and our results on evaluating summarization models in Section 6. In particular, we find that in the annual WMT evaluation for machine translation models, our methods yield the same evaluation result (model ranking) as random sampling, but with only 70% of human annotations. Given the high cost of human evaluation, this is a non-trivial cost saving.

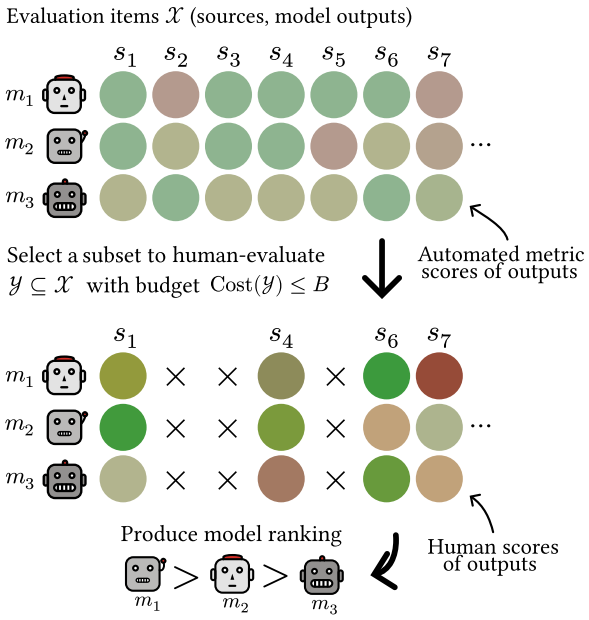

Figure 1: Output-based informative subset selection approach. Given model outputs and automated metrics, we select items to be human-evaluated on which the final model ranking can be computed.

2 Problem Statement

We are given a set of items $\mathcal{X}$ and a set of models $\mathcal{M}$ that we wish to rank according to $\mathcal{X}$. Here, each item $x \in \mathcal{X}$ is an input text and $m(x)$ is the output of the model m in the NLG task. For example, in machine translation, each x is the input in the source language and $m(x)$ is the corresponding translation. Since $|\mathcal{X}|$ is very large (exceeding the human evaluation budget B), we seek a subset $\mathcal{Y} \subseteq \mathcal{X}$ such that the ranking of models on $\mathcal{Y}$ can be as close to that on $\mathcal{X}$ as possible. An illustration of the problem is also shown in Figure 1.

Human Evaluation Set Selection. In order to obtain a subset $\mathcal{Y} \subseteq \mathcal{X}$ to be human-evaluated, we quantify the cost of human evaluations on $\mathcal{Y}$ with Cost($\mathcal{Y}$) and the usefulness of the subset for evaluation with Utility($\mathcal{Y}$). In the ideal case, the utility of the set $\mathcal{Y}$ indicates how close the human ranking of models on items $y \in \mathcal{Y}$ is to the human ranking of models on whole $\mathcal{X}$. We frame this as a subset selection problem as follows:

$$\arg\max_{\mathcal{Y} \subseteq \mathcal{X}} \text{Utility}(\mathcal{Y}) \quad \text{s.t.} \quad \text{Cost}(\mathcal{Y}) \leq B \qquad (1)$$

In our work, we make two simplifications. First, we note that the cost of evaluating a set of items is the sum of costs for evaluating individual items, i.e., $\text{Cost}(\mathcal{Y}) = \sum_{y \in \mathcal{Y}} \text{Cost}(y)$. This assumption is generally true if the human evaluation of items y is carried out one by one. Second, we assume that the utility of the set $\mathcal{Y}$ is the sum of the utilities of items $y \in \mathcal{Y}$. Note that this is generally not true, as similar datapoints may not offer an extra marginal utility for model ranking. However, in our experiments, we find that this assumption leads to good empirical results. Moreover, our preliminary experiments on modeling diversity of datapoints did not lead to better performance (see methods without this assumption in Appendix A).

With the two assumptions, we rewrite Equation (1) using 0/1 indicator variables z_x, which becomes a 0–1 knapsack problem.

$$\begin{aligned} \text{maximize} \quad & \sum_{x \in \mathcal{X}} z_x \cdot \text{Utility}(x) \\ \text{subject to} \quad & \sum_{x \in \mathcal{X}} z_x \cdot \text{Cost}(x) \leq B \\ & \text{and } \forall x \in \mathcal{X} : z_x \in \{0, 1\} \end{aligned} \qquad (2)$$

This problem is generally NP-complete (Karp et al., 1975). However, for constant Cost(.) and positive item utilities, we can find an optimal solution taking the items with the highest utilities until the budget B is reached. For a non-constant Cost(.), we use approximate integer linear programming solvers (Huangfu and Hall, 2018).

Next in Section 3, we present multiple ways to approximate the utility of an item Utility(x).

2.1 Subset Selection Evaluation

Once we select a subset of data for human evaluation, how do we know if it is good? We evaluate the quality of the selected subset with **soft pairwise accuracy** (Thompson et al., 2024) between the model ranking on the subset and the model ranking on the entire set. The accuracy reveals how close our evaluation result (model ranking) is to the true evaluation result (model ranking on the whole set). Note that this meta-evaluation requires the presence of human scores.

There are many ways to evaluate the similarity of model rankings (Deutsch et al., 2021, 2023). In contrast to Spearman (Spearman, 1904) or Kendall (Kendall, 1938) correlations, soft pairwise accuracy inherently also accounts for the statistical power of the ranking. We further discuss the choice of this meta-evaluation in Appendix B.

2.2 Source- vs. Output-based Selection

In many practical scenarios, for example, in the WMT shared tasks, the size of $\mathcal{X}$ may be so large that it is too expensive to even obtain model outputs on the entire dataset for all models. Furthermore, we may not yet know all the participating models $\mathcal{M}$ in advance, for example, when preparing blind test sets for a shared task. For such scenarios, we introduce source-based data selection. In contrast to output-based data selection, which models item utilities with the knowledge of model outputs $m(x)$ for $m \in \mathcal{M}$, source-based data selection models Utility(x) based solely on the input x.

3 Output-based Selectors

We begin by describing various utility functions for output-based subset selection to use in Equation (1). We will describe methods for source-based selection later, as these rely on the methods for output-based selection. Finally, we will compare data selection approaches using the defined utilities with a random subset selection baseline that randomly chooses a subset of data with the cost B for human evaluation. We run this random sampling 100 times and compute the corresponding confidence intervals with Student's t distribution.

3.1 Metrics Moment

We first introduce two utility functions for selecting informative items based on the distribution moments induced by automated metric scores. Metrics provide us with a coarse estimate of item difficulty, and the shape of the distribution can be used to select high-impact items for human evaluation.

Our first heuristic for defining utility is based on average metric scores. If, on average, an item receives a low metric score across multiple model outputs, then it can be perceived as being difficult (Don-Yehiya et al., 2022). Thus, average metric scores correlate negatively with the difficulty of the item for NLG models. Based on this, the first heuristic selects items with highest difficulty, i.e., lowest average metric score:

$$\textsc{MetricAvg}(x, \mathcal{M}) = - \sum_{m \in \mathcal{M}} \frac{\text{metric}(x, m(x))}{|\mathcal{M}|} \tag{3}$$

However, items where all models produce very high quality or very low quality outputs do not contribute much to the final model ranking (Zhan et al., 2021). Thus, to prioritize high-impact items that highlight differences between models, our second metric measures variance in metrics across models, which impact the final model ranking the most:

$$\textsc{MetricVar}(x, \mathcal{M}) = \sum_{m \in \mathcal{M}} \frac{(\text{metric}(x, m(x)) - \textsc{MetricAvg}(x, \mathcal{M}))^2}{|\mathcal{M}|} \tag{4}$$

3.2 Metric Consistency

Another approach to make use of the shape of the distribution of metric scores over various models is to consider items where the automated metrics show model ranking consistent with the whole set $\mathcal{X}$. Therefore, if an item is predictive of the overall model ranking based on metrics, it might also be predictive of the overall model ranking based on human scores:

$$\textsc{MetricCons}(x, \mathcal{M}) = \text{RankCorr}\left(\begin{matrix} \langle \text{metric}(x, m(x)) | m \in \mathcal{M} \rangle, \\ \langle \sum_{x' \in \mathcal{X}} \text{metric}(x', m(x')) | m \in \mathcal{M} \rangle \end{matrix} \right) \tag{5}$$

The ranking correlation might be based on Kendall, pairwise accuracy, or Spearman, which

we use. This method relates to greedy coreset construction, which we discuss in Section 7 and Appendix A.

3.3 Output Diversity

While metrics variance prioritizes items that lead to outputs of different qualities, our fourth metric goes a step further and prioritizes items that lead to diverse *outputs* as evaluating identical outputs from different models is not useful:

$$\textsc{Diversity}(x, \mathcal{M}) = \\ - \frac{\sum_{m_1, m_2 \in \mathcal{M}} \text{sim}(m_1(x), m_2(x))}{|\mathcal{M}|^2} \quad (6)$$

Output diversity can be captured by average text similarity among the outputs, for example, with pariwise unigram overlap, text-matching metrics chrF (Popović, 2015) or BLEU (Papineni et al., 2002), or embedding similarity. We primarily use embedding similarity with details in Appendix C, but also evaluate other similarity metrics in Appendix A. We take the negative value of the average to prioritize items with a lower similarity across outputs.

3.4 Item Response Theory

Previous works on informative test construction often use Item Response Theory (IRT, Santor and Ramsay, 1998), which is a principled approach to the problem. Inspired by psychometrics and educational sciences, IRT provides a way to create a budget-efficient test for evaluating and comparing various students. Given a set of students' responses to a set of test items, IRT models the probability that a student m with a given ability level θ_m will answer a question x correctly as the standard logistic function:

$$p(r_{m,x} = 1) = \frac{c_x}{1 + \exp[-a_x(\theta_m - b_x)]} \quad (7)$$

Here, the student response (r) is usually a binary label (correct or incorrect), a_x denotes the **discriminability** of the item, b_x denotes the **difficulty** of the item x, and c_x denotes the maximum achievable chance of answering the items (e.g., due to ambiguity of the item). The item parameters define the shape of the standard logistic function: For items with high discriminability a_x, even very close students can be distinguished because a small change in θ_m changes the prediction.

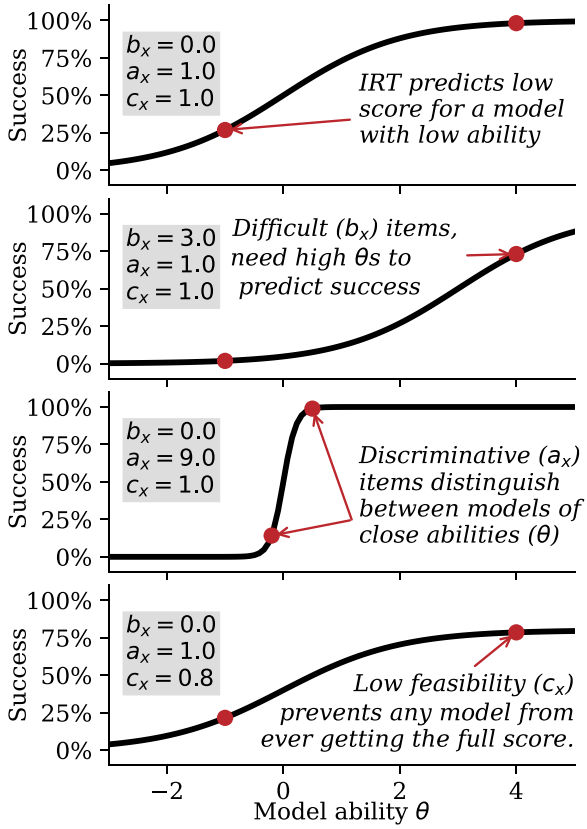

Figure 2: Illustration on how item response theory predicts model success. The parameters are difficulty b_s (shift on x-axis), discriminability a_s (slope), and feasibility c_s (upper bound).

We draw on this analogy and fit an IRT model to predict the metric scores $\text{metric}(x, m(x))$ for various models m on the items x. However, in NLG, the metric scores are usually not binary, but continuous. Some approaches bypass this by binarizing the continuous variable: $r_{m,x} := \mathbb{1}[\text{metric}(x, m(x)) > 0.5]$ (Polo et al., 2024). However, this leads to a loss of information, and we can redefine the objective to directly predicting the continuous score $r_{m,x} := \text{metric}(x, m(x))$ as in other works in psychometry (Noel and Dauvier, 2007):

$$\hat{r}_{m,x} = \frac{c_x}{1 + \exp[-a_x(\theta_m - b_x)]}, \quad (8)$$

See an illustration of IRT predictions and parameters b_x, a_x, and c_x in Figure 2. Following standard practice, we optimize the IRT model with stochastic variational inference (Wu et al., 2020; Rodriguez et al., 2021; Lalor and Rodriguez, 2023) using the ELBO loss. In practice, this corresponds to making the estimated response $\hat{r}_{m,x}$ be closer to the true response $r_{m,x}$.

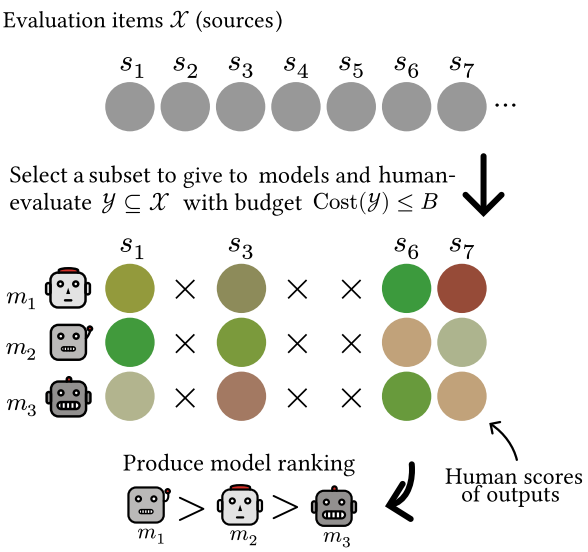

Figure 3: Source-based variant of subset selection (in contrast with output-based in Figure 1). Given just the item inputs, we select items to be given to models and their outputs to be human-evaluated.

We repurpose this method from evaluating students to evaluating models and use both the item discriminability a_x, and difficulty b_x for utility:

$$\textsc{DiffDisc}\,(x) = a_x \times b_x \tag{9}$$

We use a product instead of addition because a_x and b_x may have different scales but consider also other formulas in Appendix A.

4 Source-based Selectors

Next, we discuss approaches for source-based data selection where the model outputs and metric scores are not available yet. This is illustrated in Figure 3. We provide two ways to adapt our output-based methods to the case where we select datapoints given only source texts: (1) estimating item utility just based on x via **model distillation**, and (2) creating an **artificial crowd** of models imitating $\mathcal{M}$.

4.1 Item Utility Distillation

Calculating item utilities in output-based selection requires access to the outputs of models $\mathcal{M}$. To circumvent this requirement in source-based selection, we develop a model distillation approach to fit a model that predicts the item utility in the absence of $\mathcal{M}$, just based on the item input x.

$$\text{Utility}(x, \mathcal{M}) \xrightarrow{\text{distill}} \text{Utility}^{\text{src}}(x) \tag{10}$$

We use the architecture of a learned NLG metric (Rei et al., 2020) to predict the utilities. The input text x is encoded with a pre-trained language model, and a regression head is trained with MSE loss predicting output-based utilities. As a result, we obtain new source-based item utility estimators: $\textsc{MetricAvg}^{\text{src}}$, $\textsc{MetricVar}^{\text{src}}$, $\textsc{MetricCons}^{\text{src}}$, $\textsc{MetricDiversity}^{\text{src}}$, and $\textsc{DiffDisc}^{\text{src}}$. For IRT, Benedetto et al. (2020) and Byrd and Srivastava (2022) build functions that directly predict the item discriminability a_x and difficulty b_x for unseen items. In contrast, our predictor directly predicts the product $a_x \times b_x$ used for computing the item utility.

An advantage of source-based item utilities over output-based ones is that we can train the model to predict the utilities based on human scores as opposed to metric scores, if these exist in the training data. In the context of machine translation, similar models are used to estimate the properties of translations and sources (Rei et al., 2020; Don-Yehiya et al., 2022; Zouhar et al., 2023; Perrella et al., 2024).

See implementation details in Appendix C. Note that this method only works on novel items that are similar to those in the training data.

4.2 Artificial Crowd

In some cases, we might not know the complete set of models $\mathcal{M}$ or not have the computational capacity to compute the output for all the models in $\mathcal{M}$. However, we might still have some prior knowledge of $\mathcal{M}$, such as knowing which particular language models will be present. We now assume that we know $\mathcal{M}'$ such that $\mathcal{M}' \approx \mathcal{M}$ in some capacity. Then we can use the outputs from the models $\mathcal{M}'$ to approximate the output-based utilities without exact knowledge of $\mathcal{M}$:

$$\textsc{Utility}(x, \mathcal{M}') \approx \textsc{Utility}(x, \mathcal{M}) \tag{11}$$

In our implementation, we take a subset of the original $\mathcal{M}$, simulating the case where we are willing to compute model outputs for at least a portion of the models. The approach of designing an artificial crowd was previously used by Lalor et al. (2019) to train IRT in sentiment classification and natural language inference tasks.

5 Case Study 1: Machine Translation

As our first case study for budget-efficient human evaluation, we consider machine translation,

particularly the setting in the general WMT shared task (Kocmi et al., 2024a). For WMT and similar venues, both source-based and output-based variations come into play sequentially: (1) all source data is collected, (2) an initial subset is created based on just sources, (3) the initial subset is distributed to participants & outputs are collected, (4) automated metrics are computed, (5) the final subset is created based on sources and outputs, (6) the final subset is human-evaluated, and (7) the model ranking is produced.

We first describe the experimental setup by which we evaluate our data selection methods. We then present the results for the simpler of the two scenarios: output-based selection, considering both segment-level translation and document-level translation (Section 5.1). Then, we explore the case considering the annotation cost of individual items and their impact on subset selection (Section 5.2). We conclude with the results for the source-based data selection task (Section 5.3). In our results, we focus on selecting data subsets of sizes ranging from 5% to 50% of the original test set.

In order to satisfy the positivity constraints of the item utility functions (Section 2), we shift all the utilities to be positive by a constant. In our initial experiments, we assume a constant human evaluation cost for each item.

Data Setup. For our experiments, we use the human annotation data from publicly available past WMT campaigns. We include only the language pairs and WMT years with at least 500 human annotations per model. As a result, we ended with 33 campaigns with 31k source items and 395k translations. Of these, we use the nine language pairs from WMT 2023 that contain MetricX-23 (Czech→Ukrainian, German→English, Japanese→English, Chinese→English, English→Czech, English→German, English→Japanese, English→Chinese, Hebrew→English). Unless specified otherwise, we average the results (soft pairwise accuracy) across all languages.

5.1 Output-based Selection for WMT

We evaluated the efficacy of our proposed methods using soft pairwise accuracy (Section 2.1). The results of output-based selection (that is, when models and metric scores are known) are shown in Figure 4 at specific subset sizes. We compare these methods with random selection, which has

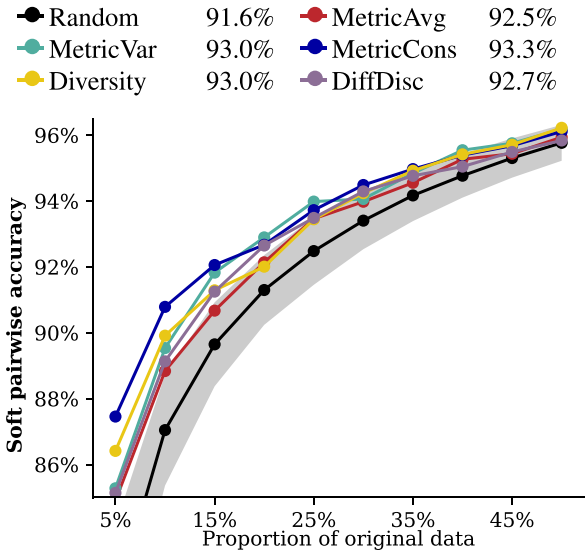

Figure 4: Main **output-based results for machine translation** (WMT23) with soft pairwise accuracy. Subset selection methods are based either on MetricX-23 or output diversity. We also show 90% t-distribution confidence intervals for the Random Selector from 100 runs. Numbers in the legend show average soft pairwise accuracy across all data proportions.

been reported to be a strong baseline in previous works (Rodriguez et al., 2021; Park et al., 2022; Polo et al., 2024). We discuss the reason for random being a strong baseline in Appendix B.

We find that approaches based on metrics moments (average and variance of automated metrics) surpass the random baseline confidence interval only with a slight margin when measured by soft pairwise accuracy. Additionally, the IRT model performs slightly better than the metrics moment approaches, and diversity performs best. Note that even improvements within the random baseline confidence interval but above the mean are meaningful, because by sampling randomly, we are likely to arrive at a subset with soft pairwise accuracy anywhere within the shaded area, so worse than the methods. The results across individual WMT years and languages are shown in the Appendix Table 10. In Appendix B we further confirm that even when the subset selection methods are meta-evaluated differently from soft pairwise accuracy (e.g., by correlations between model rankings), they still consistently outperform random selection.

We now cover two practical aspects of dataset creation: (1) selecting with some apriori item distribution, such as by having a specific number

Method	Output-based	Source-based$^{(src)}$
Random	91.3%$_{\pm 0.30\%}$	91.3%$_{\pm 0.30\%}$
MetricAvg	92.4%	91.8%
MetricVar	92.6%	91.9%
MetricCons	93.2%	93.0%
Diversity	92.9%	92.6%
DiffDisc	92.4%	92.1%

Table 1: Subset selection with balanced domains. The top-$^{B}/|\mathcal{D}|$ of item utilities within each domain in $\mathcal{D}$ is taken for particular budget B. $\pm$ shows 90% t-test confidence interval from 100 runs. Results are averaged across languages and subset sizes.

Method	Output-based	Source-based$^{(src)}$
Random	87.2%$_{\pm 0.28\%}$	87.2%$_{\pm 0.28\%}$
MetricAvg	90.5%	88.2%
MetricVar	89.5%	88.7%
MetricCons	89.3%	89.2%
Diversity	89.6%	89.0%
DiffDisc	90.1%	89.2%

Table 2: Document-level subset selection. The item utilities are averaged to create document-level utilities out of which a subset is chosen. $\pm$ shows 90% t-test confidence interval from 100 runs. Results are averaged across languages and subset sizes.

of items in each data domain, and (2) selecting higher-level item units, such as documents, while items correspond to paragraphs within the document.

Balancing Domains. Items in the WMT data come from multiple distinct domains, such as news, social, or chat. As a result, the subset would measure fewer ability directions, which would lead to more significance when comparing models. To test this, we enforce equal number of selected items in each domain (Table 1). The results in Table 2 (left column) show that even with this balancing, our methods perform much better than random selection.

Selecting Document-level Items. Recently, MT evaluation has shifted to document-level translation evaluations (Kocmi et al., 2024a, inter alia) where entire document translations are now evaluated. We extend our methods to this setting by defining the utility of a document as the average of various sentence-level utilities defined in this paper. We now consider entire documents to be items where the cost and utility of the document are the sum of costs of evaluating sentence-level translations in the document and average of sentence-level utilities, respectively. The results in Table 2 mimic the main results for item-level selection and outperform random selection.

Other Methods. In Appendix A we describe and evaluate other subset selection methods, spanning baselines, oracles, and related work. Notably, apart from oracles, no method consistently outperforms the methods mentioned in the main paper.

5.2 Accounting for Human Annotation Cost in Subset Selection

The cost of human evaluation of model outputs is usually determined by the total time that human annotators spend evaluating them. However, different annotation items take different times to annotate due to their lengths. For example, the machine translation evaluation campaign of Kocmi et al. (2024c) contains very short 3-word items that take less than 10 seconds to annotate but also multi-sentence items that take more than 3 minutes to annotate each.

As human evaluation time is not known during the subset selection stage, we use source length as a rough measure for human evaluation time. Kocmi et al. (2024c) provide a human evaluation of the machine translation dataset with evaluation times. We used this data to approximate the human evaluation time as a linear function of source length. $\hat{t}(x) = 0.15 \cdot |x| + 33.7$. This approximation is weakly correlated with the real human evaluation time ($\rho = 0.24$). Furthermore, items with the highest difficulty (lowest metric scores, MetricAvg) correlate positively ($\rho = 0.33$) with human evaluation time. Therefore, an item difficult for models usually has a higher annotation cost compared to the rest of the test set. Selecting the most difficult items to human evaluate may thus be suboptimal and lead to higher costs.

To avoid accidentally increasing the cost of the evaluation, we need to take the cost into account during subset selection. For this, we approximate the solution to the optimization in Equation (2) with relaxed integer linear programming. The results in Table 3 show that cost-aware subset selection can lead to a very high soft pairwise accuracy even with a limited budget.

Method	Output-based	Source-based (src)
Random	$93.9\%^{\Delta 2.3\%}_{\pm 0.38\%}$	$93.9\%^{\Delta 2.3\%}_{\pm 0.38\%}$
MetricAvg	$94.6\%^{\Delta 2.1\%}$	$93.7\%^{\Delta 1.2\%}$
MetricVar	$94.8\%^{\Delta 1.8\%}$	$93.9\%^{\Delta 0.9\%}$
MetricCons	$94.5\%^{\Delta 1.2\%}$	$94.3\%^{\Delta 1.0\%}$
Diversity	$94.4\%^{\Delta 1.4\%}$	$94.2\%^{\Delta 1.2\%}$
Diff.×Disc.	$94.7\%^{\Delta 1.9\%}$	$94.2\%^{\Delta 1.5\%}$

Table 3: Results for methods with cost-aware subset selection with integer linear programming (Equation 2). The sizes of the resulting subsets can differ but have the same cost. Δ indicates improvements against cost-unaware selection in Figure 4. $\pm$ shows 90% t-test confidence interval from 100 runs. Results are averaged across languages and subset sizes.

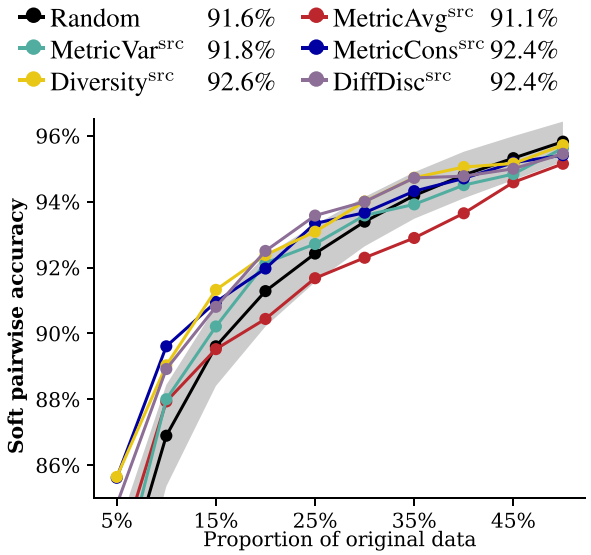

Figure 5: Main **source-based results for machine translation** (WMT23) with soft pairwise accuracy. Utility predictors are based either on distilling human scores or output diversity. We also show 90% t-distribution confidence intervals for the Random Selector from 100 runs. Numbers in the legend show average soft pairwise accuracy across all data proportions.

5.3 Source-based Selection for WMT

We now describe the source-based selection case where we do not yet know the model outputs for $\mathcal{X}$. The first approach is to distill the item utilities with a source-based predictor.

We train all item utility estimator models by distilling utilities (Equation 10) only up to WMT22 to avoid contamination of the evaluation on WMT23. Figure 5 shows the source-based selection results, which can be applied to unseen items without model outputs, unlike output-based selection. Although the source-based estimation of metric moments performs worse than in output-based selection (Figure 4), it still outperforms random selection, although not outside the confidence interval. Diversity$^{\text{src}}$, which predicts output diversity, and Diff.×Disc.$^{\text{src}}$, which predict the latent IRT parameters, perform the best, though still falling short of their output-based selection versions.

Method	Artificial Crowd	Source-based(src)
Random	$91.8\%_{\pm 0.89\%}$	$91.8\%_{\pm 0.89\%}$
MetricAvg	92.0%	90.7%
MetricVar	92.5%	92.1%
MetricCons	92.3%	92.1%
Diversity	92.6%	92.5%
DiffDisc	91.9%	92.4%

Table 4: Output-based selection methods applied to source-based selection using artificial crowd of size of 4 models, randomly sampled from $\mathcal{M}$. The results differ from source-based selection (Figure 5) because the selection is evaluated on $\mathcal{M} \setminus \mathcal{M}'$. $\pm$ shows 90% t-test confidence interval from 100 runs. Averaged across languages and subset sizes.

For the artificial crowd approach, we assume that we know the model output on $\mathcal{X}$ for at least some models. However, this does not require past evaluation data. Table 4 shows that the previous distillation methods for source-based selection perform on par with the artificial crowd approach. The choice whether to use an artificial crowd or item utility predictors for the source-based subset selection case then depends on availability of at least some model outputs or past evaluation data on which an estimator can be trained.

Unfair Evaluation Bias. Using the artificial crowd might lead to a biased test set construction when used together with MetricAvg. If a model m is being evaluated ($m \in \mathcal{M}$) and is also part of the artificial crowd ($m \in \mathcal{M}'$), then m is at a disadvantage because we selected difficult items for m but not necessarily for $\mathcal{M} \setminus \{m\}$. This can also happen distributionally on a higher level when we use an artificial crowd consisting of some kind of models, such as multilingual language models, which will then make the test set more difficult to multilingual language model specifically.

Method	Output-based	Source-based (src)
~~Random~~	~~100.0%~~	~~100.0%~~
MetricAvg	85.0%	106.9%
MetricVar	81.0%	97.2%
MetricCons	⋆ 71.4%	90.1%
Diversity	76.8%	⋆ 78.7%
Diff.×Disc.	87.2%	91.5%

Table 5: Proportion of data needed to reach the same evaluation result for WMT23 (soft pairwise accuracy) as random subset selection. Averaged across budgets from Figures 4 and 5.

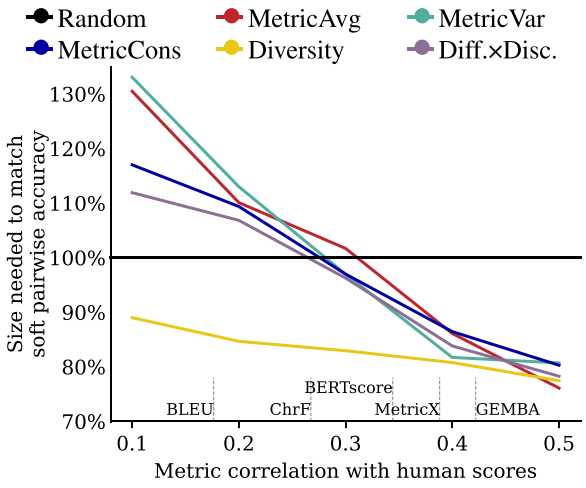

Figure 6: Proportion of data needed to reach the same evaluation result for WMT23 (soft pairwise accuracy) with respect to the automated metric used for informing the selection. The automated metric correlation quality (x-axis) is measured as item-level Pearson correlation against human scores. Each value is a separate dataset and metric, averaged in bins 0.1, 0.2, 0.3, 0.4, 0.5. Even though the diversity method is invariant to the metrics, it is affected by the particular datasets characteristics (e.g., quality of human annotations), which differ across bins.

The same problem extends to MetricAvg$^{\text{src}}$, or any other learned metrics, when we consider that they were trained on human quality assessments of some model outputs $\mathcal{M}_{\text{prev}}$. Again, if a model m that is being evaluated $m \in \mathcal{M}$ is also $m \in \mathcal{M}_{\text{prev}}$, then m is again at a disadvantage.

Together with the potential of selecting costly-to-evaluate items, we advise against MetricAvg for subset selection for both output- and source-based variants, especially when using an artificial crowd.

5.4 Subset Selection Cuts Costs Substantially

At first glance, the improvements over random selection in both scenarios appear minor. Although previous work on active learning and subset selection agrees that random sampling is a strong baseline (Wei et al., 2015; Rodriguez et al., 2021; Park et al., 2022; Polo et al., 2024), the improvements shown do matter, when applied at a larger scale. To show this, we ask: *What budget would we need to arrive at the same evaluation result as random sampling with budget B?* The answer can be obtained with the following formula:

$$\hat{C} = \min\{C | \text{SPA}(\mathcal{Y}^{\dagger}_{\leq C}) \geq \text{SPA}(\mathcal{Y}^{R}_{\leq B})\} \quad (12)$$

where SPA is soft pairwise accuracy, $\mathcal{Y}^{\dagger}_{\leq C}$ is a subset within budget C that is comparable to a random subset $\mathcal{Y}^{R}_{\leq B}$ of budget B.

In Table 5, we quantify the cost-efficiency of our subset selection approach by reporting the proportion of number of datapoints in our subset that achieves the same evaluation result as random sampling, $\hat{C}/B$. We find that even the simple diversity–based utility approach achieves the same or better soft pairwise accuracy with only 77% of the data as random sampling on the machine translation evaluation task. This means that to achieve the same human evaluation quality, we need to now pay for evaluation of only a portion of the test data. At the scale of industrial evaluation of NLP models, which gets bigger each year, these economic implications are substantial.

5.5 Importance of a Good Automated Metric

In our main results for machine translation (Figure 4), we used MetricX-23, which is one of the best automated machine translation metrics (Freitag et al., 2023). *What happens with a weaker metric?* In Figure 6, we show the relationship between evaluation subset selection performance and quality of the metric, as measured by item-level correlations with human evaluations. We show results across the 25 metrics available in the WMT metrics shared task (Freitag et al., 2021, 2022, 2023).

MetricX-23 is a pre-trained supervised metric that requires human evaluation data for training. BERTscore (Zhang et al., 2020) is also a pre-trained metric, based on a language model, but is not fine-tuned on human data. BLEU (Papineni et al., 2002) and chrF (Popović, 2015) are string matching metrics that do not need pre-training or

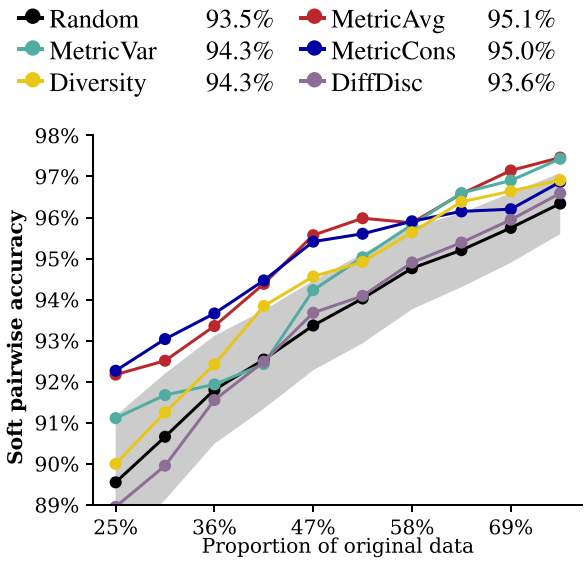

Figure 7: Main **output-based results for SummEval** averaged across relevance, coherence, consistency, fluency, and their sum with soft pairwise accuracy. The utility predictors are metrics moments (MetricAvg, MetricVar), consistency (MetricCons) and Item Response Theory (DiffDisc) based on G-Eval, and output diversity. Numbers in legend show averages across all points. We also show 90% t-distribution confidence intervals for the Random Selector from 100 runs. Numbers in the legend show average soft pairwise accuracy across all data proportions.

data. Lastly, GEMBA (Kocmi and Federmann, 2023) is an LLM-as-a-Judge approach to an automated machine translation metric based on GPT-4.

The string matching metrics (BLEU, chrF) do not provide enough signal for robust evaluation subset selection (worse or on par with random). However, pre-trained but not supervised metrics (BERTScore) are sufficient because they correlate well enough with evaluations. Generally, the higher the quality of the automated metric, the better the evaluation subset quality. However, combining multiple automated metrics into a single item utility does not provide improvements, as shown in Appendix A. Probably, this is because a simple combination of automated metrics rarely leads to higher correlations with human judgments.

6 Case Study 2: Summarization

We now show the applicability of our methods to another natural language generation task, summarization. The summarization task is more open-ended than machine translation, making the evaluation even more difficult. In this section, we explore budget-efficient subset selection for both human evaluation and cheaper, yet still not free, large language model-based evaluation.

Method	Human	G-Eval
~~Random~~	~~100.0%~~	~~100.0%~~
MetricAvg	73.7%	84.0%
MetricVar	82.8%	84.7%
MetricCons	70.6%	87.5%
Diversity	88.7%	89.2%
DiffDisc	99.5%	78.6%

Table 6: Proportion of data needed to reach the same evaluation result on SummEval (soft pairwise accuracy) as random subset selection with respect to human on LLM evaluation. Averaged across Figure 7 budgets.

Setup. We use SummEval (Fabbri et al., 2021) which contains 17 human-evaluated models on 100 items. The human evaluation of each model output, in contrast to machine translation, is not a single scalar for overall quality. Instead, humans evaluate the output in four dimensions: *relevance*, *coherence*, *consistency*, and *fluency*. We use the subset selected for evaluation on each quality dimension independently and then aggregate the final results. We choose G-Eval (Liu et al., 2023), an LLM-based evaluator, as it has a high correlation between metric predictions and human scores.

Results. The results of output-based data selection for summarization are shown in Figure 7. We do not include source-based selection because that requires training data, which are not available for a dataset this small. In most cases, the methods outperform the random selection. Similarly to machine translation, the metric average is not consistently very good. Metric variance, consistency, and diversity are again much stronger. The results in Table 6 show that we only need $\sim 70\%$ of the budget to obtain the same evaluation result as random selection when using the metric consistency selector. We show the results across the individual evaluation dimensions (relevance, coherence, consistency, and fluency) in Appendix Table 10.

Even when using the LLM-based metric for final evaluation, it can become expensive on a larger scale. Budget-efficient subset selection can

also make even this type of evaluation more economical. The results in the right column of Table 6 show that we can again select only $\sim 80\%$ of the test set to be LLM evaluated to reach the same evaluation result as random selection.

7 Related Work

In this section, we provide context for natural language generation evaluation and an overview of previous works on budget-efficient evaluation.

NLG Evaluation. Machine translation is one of the most important NLG tasks, and the progress of MT is tracked annually in various WMT evaluations (Kocmi et al., 2024a; Nakazawa and Goto, 2024; Ahmad et al., 2024). For some NLG tasks, the output is a string in natural language, and its quality can not always be easily assessed by comparing to a ground truth. This is because there might be many acceptable outputs, such as multiple equally correct translations of a single sentence. For this reason, robust NLG evaluation eventually relies on human annotators. However, the human evaluation process is also not straightforward, ranging from assigning a single score (Graham et al., 2015; Kocmi et al., 2022) to marking error spans (Lommel et al., 2014; Kocmi et al., 2024c). Typically, the output of this process is a single number per input+output, which is used to determine the final model ranking.

At scale, such as during model development, human evaluation is too costly. For this reason, automated metrics have been developed, starting from text matching approaches (Papineni et al., 2002; Popović, 2015), progressing to learned metrics (Rei et al., 2020; Juraska et al., 2024). Although learned metrics correlate more strongly with humans (Freitag et al., 2022), they also have unexpected problems (Zouhar et al., 2024b,a; Falcão et al., 2024). Like human annotations, automated metrics also produce a single number for each input+output, evaluating the quality of the NLG model's output.

Datapoint selection already begins to play a role for automated metric meta-evaluation. For example, for summarization metrics evaluation, Deutsch et al. (2022) redefine high quality metric to be those being able to distinguish between closely-performing models. To this end, the metrics performance is not meta-evaluated with correlation across the whole set (even if human scores are available), but on a subset of closely-matched systems.

Budget-efficient Evaluation. Previous work attempted to find the most informative or diverse evaluation items, primarily by comparison to a ground-truth answer. In contrast, our goal is to reduce the number of *human* evaluations, while still arriving at the same result, such as model ranking.

To select the most informative items for evaluation, Rodriguez et al. (2021) and Polo et al. (2024) use Item Response Theory (IRT, Santor and Ramsay, 1998), a framework for the creation of test sets for students in classrooms. However, these methods require ground-truth answers to which the model output is compared with a binary outcome. For natural language generation tasks, there is no single ground-truth answer, and the model output evaluation is a continuous score.

In a more complex version of stratified sampling (Saldías Fuentes et al., 2022), Ni et al. (2024) supersample items to mimic the distribution throughout the test set. The subset is constructed based on most difficult items, which again requires a comparison to a ground truth and a previous evaluation of some models on the whole test set. For the evaluation of classification models, Vivek et al. (2024) find anchor points that describe the outcome of the evaluation on the whole test set. Feng et al. (2024) choose items that have the least similar model output to be evaluated via pairwise comparisons, which is not applicable to direct model output evaluation.

Many previous works require that at least a few models have already been evaluated on the entire set of items from which we select a subset (Rodriguez et al., 2021; Feng et al., 2024; Ruan et al., 2024). This makes the methods not applicable for source-based selection, where either the evaluation set is too large to be processed or where the to-be-evaluated models are not known in advance.

Subset selection through batched active learning (Park et al., 2022; Mendonça et al., 2021, 2023; Ruan et al., 2024; Li et al., 2024) is another approach using iterative selection. In practice, human evaluations are usually run all at once or outsourced to a third party with annotators working asynchronously. Therefore, active learning is possible only when the annotation process is tightly controlled. In contrast, our item utilities

Work	Human-eval.	Scores	Output-based	Source-based	Cost-aware
Ours	✓ yes	✓ continuous	✓ yes	✓ yes	✓ yes
Rodriguez et al. (2021)	✗ ground-truth	✗ binary	✓ yes	✗ no	✗ no
Ni et al. (2024)	✗ ground-truth	✗ binary	✗ no	✓ yes	✗ no
Polo et al. (2024)	✗ ground-truth	/ continuous*	✓ yes	/ partly	✗ no
Vivek et al. (2024)	✗ ground-truth	✗ binary	✓ yes	✓ yes	✗ no
Feng et al. (2024)	✓ yes	✗ pairwise	✓ yes	✗ no	✗ no
Ashury Tahan et al. (2024)	✓ yes	✗ pairwise	✓ yes	✗ no	✗ no
Ruan et al. (2024)	✓ yes	✓ continuous	/ iterative	✗ no	✗ no
Li et al. (2025)	✗ no	✗ binary	✓ yet	✗ no	✗ no

Table 7: Prior work on budget-efficient evaluation subset selection. Ground truth: needs comparison to ground truth, scores: binary or continuous outcomes, or pairwise comparisons, output/source-based: methods for the two subset selection variants, cost-aware: can take evaluation cost into account.

can be simply used to sort the items from most to least informative, and annotators can then stop ad-hoc when the budget is reached.

Zouhar et al. (2025a) already human-evaluate only a subset of the whole test set by skipping items where automated metrics report no errors. However, this subset selection is ad hoc and does not offer any control over the subset size.

To our knowledge, no prior work has studied subset selection for evaluation in which the item evaluation cost is taken into account. See Table 7 for a high-level comparison with previous work.

Coresets. Discovering a coreset of the data is another potential approach. A coreset of a dataset is defined as a small subset such that solving a problem on the coreset yields the same result as solving the same problem on the original set (Jubran et al., 2019). Although appearing like the golden bullet for our problem, it is not applicable for two reasons. First, we do not know the information (human scores) of the full set that we are trying to reduce. Second, the coreset algorithms make use of properties of the loss (subset evaluation) function. Although optimization algorithms exist to, for example, find a subset of vectors with the same mean (approximate mean coreset, Vahidian et al., 2020), to our knowledge, none exist to optimize soft pairwise accuracy or other related ranking evaluation. Finally, coreset algorithms are often about reducing computational costs in subset construction, which is not the case in our setup. In Appendix A we examine a brute-force imitation of coreset construction and show that due to the first point (lack of human scores at the time of selection), these methods fall short.

8 Conclusion and Key Takeaways

We formalize the task of selecting subsets from the test set with the goal of selecting a subset of the test set for efficient human evaluation. We explore two common variants of this task: source-based selection (no model outputs available), and output-based selection (outputs and automated metrics available). We present several methods based on metric variance, consistency, and diversity; and show that they outperform the dominantly used random selection approach for machine translation (Section 5) and summarization evaluations (Section 6). However, the simple heuristic of using difficulty estimate by prioritizing items with lowest automated metric scores does not lead to large improvements and can be counterproductive as it prioritizes costly-to-annotate items. All our methods are implemented in the subset2evaluate package with pre-trained item utility estimators ready to use for subset selection in machine translation.

Takeaways. Based on the analysis in this paper, we offer the following advice to NLG researchers and practitioners for selecting a subset of data for human evaluation.

- If model outputs and reliable metrics are available, use metric variance or metric consistency.
- If model outputs are available but a reliable metric is not available, use output diversity.
- If model outputs are not available but historical data are available, use the diversity$^{\text{src}}$ estimation.
- If only some model outputs are available, use an artificial crowd with metric variance.

Limitations. A key limitation of our approach is the potential bias of automated metrics and artificial crowd selection. Automated metrics may already have been used to inform evaluation procedures, although they are misaligned with the

final human judgments (Kocmi et al., 2024b,a). Data selection for human evaluation would make it harder to catch these biases. To illustrate this issue, consider a model that underperforms on specific types of item that is not detected by the automated metric. As these items are estimated to not be faulty, they are not selected for human evaluation. This issue mainly arises when using expected metric averages for item selection, although similar challenges could occur with variance- or discrimination-based methods (see discussion in Section 5.3). Section 5.5 shows that the performance of our methods directly corresponds to the alignment between automated metrics and human judgments. Thus, our methods would continue to benefit from future improvements in automated metrics.

Ethical Considerations. We reuse existing data from WMT and SummEval and do not employ our own annotators. In the context of automatization and job security, our work aims not to substitute human work, but to ensure that the effort is not wasted on annotating less informative items, ultimately making the work more meaningful.

Acknowledgments

We thank the TACL reviewers and action editor for a productive reviewing process. This research has been funded in part by a Swiss National Science Foundation award (project 201009) and a Responsible AI grant by the Haslerstiftung.

References

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, and Rodolfo Zevallos. 2024. Findings of the IWSLT 2024 evaluation campaign. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Shir Ashury Tahan, Ariel Gera, Benjamin Sznajder, Leshem Choshen, Liat Ein-Dor, and Eyal Shnarch. 2024. Label-efficient model selection for text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8384–8402. Association for Computational Linguistics.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. R2DE: A NLP approach to estimating IRT parameters of newly generated questions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 412–421.

Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146. https://doi.org/10.1162/tacl_a_00417

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, pages 6038–6052. Association for Computational Linguistics.

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

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2022. PreQuEL: Quality estimation of machine translation outputs in advance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11183. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409. https://doi.org/10.1162/tacl_a_00373

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

Kehua Feng, Keyan Ding, Kede Ma, Zhihua Wang, Qiang Zhang, and Huajun Chen. 2024. Sample-efficient human evaluation of large language models via maximum discrepancy competition.

Ronald A. Fisher. 1935. *The Design of Experiments*. Oliver & Boyd, Edinburgh.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774. Association for Computational Linguistics.

Phillip Good. 2000. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer. https://doi.org/10.1007/978-1-4757-3235-1

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191. Association for Computational Linguistics.

Q. Huangfu and J. A. J. Hall. 2018. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142.

Ibrahim Jubran, Alaa Maalouf, and Dan Feldman. 2019. Introduction to coresets: Accurate coresets.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the WMT 2024 metrics shared task.

Richard M. Karp, Raymond E. Miller, and James W. Thatcher. 1975. Reducibility among combinatorial problems. *Journal of Symbolic Logic*, 40(4).

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich,

Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. 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, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024b. Preliminary WMT24 ranking of general MT systems and LLMs.

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

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4.In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024c. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.

John Patrick Lalor and Pedro Rodriguez. 2023. `py-irt`: A scalable item response theory library for python. *INFORMS Journal on Computing*, 35(1):5–13. `https://doi.org/10.1287/ijoc.2022.1250`

John Patrick Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2019, page 4240. NIH Public Access.

Xiang Lisa Li, Farzaan Kaiyom, Evan Zheran Liu, Yifan Mai, Percy Liang, and Tatsunori Hashimoto. 2025. AutoBencher: Towards declarative benchmark construction.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. 2024. Active evaluation acquisition for efficient LLM benchmarking.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:0455–463.

Frederic M. Lord and Melvin R. Novick. 2008. *Statistical theories of mental test scores*. IAP.

Vânia Mendonça, Ricardo Rei, Luísa Coheur, and Alberto Sardinha. 2023. Onception: Active learning with expert advice for real world machine translation. *Computational Linguistics*, 49(2):325–372. `https://doi.org/10.1162/coli_a_00473`

Vânia Mendonça, Ricardo Rei, Luisa Coheur, Alberto Sardinha, and Ana Lúcia Santos. 2021. Online Learning meets Machine Translation evaluation: Finding the best systems with the least human effort. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3105–3117. Association for Computational Linguistics.

Toshiaki Nakazawa and Isao Goto, editors. 2024. Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wat-1.0

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. MixEval: Deriving wisdom of the crowd from LLM benchmark mixtures.

Yvonnick Noel and Bruno Dauvier. 2007. A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, 31(1):47–73.

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

Dongmin Park, Dimitris Papailiopoulos, and Kangwook Lee. 2022. Active learning is a strong baseline for data subset selection. In *Has it Trained Yet? NeurIPS 2022 Workshop*.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244. Association for Computational Linguistics.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinyBenchmarks: Evaluating LLMs with fewer examples.

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

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503. Association for Computational Linguistics.

Jie Ruan, Xiao Pu, Mingqi Gao, Xiaojun Wan, and Yuesheng Zhu. 2024. Better than random: Reliable NLG human evaluation with constrained active sampling.

Belén Saldías Fuentes, George Foster, Markus Freitag, and Qijun Tan. 2022. Toward more effective human evaluation for machine translation. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 76–89. Association for Computational Linguistics.

Darcy A. Santor and James O. Ramsay. 1998. Progress in the technology of measurement: Applications of item response models. *Psychological Assessment*, 10(4):345.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

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

Saeed Vahidian, Baharan Mirzasoleiman, and Alexander Cloninger. 2020. Coresets for estimating means and mean square error with limited greedy samples. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 350–359. PMLR.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2024. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1601. Association for Computational Linguistics.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2015. Submodularity in data subset selection and active learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1954–1963. Lille, France. PMLR.

Mike Wu, Richard L. Davis, Benjamin W. Domingue, Chris Piech, and Noah Goodman. 2020. Variational item response theory: Fast, accurate, and expressive.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021. Variance-aware machine translation test sets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324. Association for Computational Linguistics.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288. Association for Computational Linguistics.

Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang, and Mrinmaya Sachan. 2023. Poor man's quality estimation: Predicting reference-based MT metrics without the reference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1311–1325. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500. Association for Computational Linguistics.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025a. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

Vilém Zouhar, Maike Züfle, Beni Egressy, Julius Cheng, and Jan Niehues. 2025b. Early-exit and instant confidence translation quality estimation.

A Other Methods

We now describe and evaluate a collection of subset selection methods which are not included in the main paper either because of their assumptions or low performance, which limits their application.

A.1 Methods

For simplicity and comparability across methods that do not support cost-aware subset selection, we assume constant cost of each item in this section. Thus we assume $\text{Cost}(\mathcal{Y}) = |\mathcal{Y}|$ in the optimization objective from Equation (1).

Bruteforce Subset Selection. For the bruteforce approach we simply sample N subsets $\hat{\mathcal{Y}}$ of size $|\hat{\mathcal{Y}}| = B$ and pick the one with the highest utility Utility($\hat{\mathcal{Y}}$). Evaluating Utility($\mathcal{Y}$) requires computing soft pairwise accuracy with respect to the full set $\mathcal{X}$ using human scores, to which we do not have access at the time of subset creation. However, we can either consider the human scores as an oracle, or compute soft pairwise accuracy with respect to the proxy metric, such as MetricX-23. Sampling only 1 subset corresponds to the random baseline. Importantly, this optimization does not require the second simplification put forth in Section 2, namely that items contribute value to the evaluation independently and not jointly.

Greedy Subset Selection. The previous method relies solely on the variance of random selection and taking the maximum (as in Figure 4). A natural extension is to start with a small random set and iteratively extend it by examples that increase the soft pairwise accuracy the most. Specifically, we keep $\hat{\mathcal{Y}}_i$ at each step i and choose the extension with size S:

$$\hat{\mathcal{Y}}_{i+1} = \hat{\mathcal{Y}}_{i+1} \cup \underset{\substack{\mathcal{Y}^+ \subseteq \mathcal{X} \setminus \hat{\mathcal{Y}}_i \\ |\mathcal{Y}^+| = S}}{\arg\max}\ \text{Utility}(\hat{\mathcal{Y}}_\rangle \cup \mathcal{Y}^+) \quad (13)$$

We stop iterating steps when $|\hat{\mathcal{Y}}_i| = B$. The bruteforce subset selection is a special case where $S = B$. The greedy subset selection is also a special case of beam search, with beam size of 1.

Human Oracle. The existing methods, such as greedy subset selection, MetricAvg, MetricVar, and MetricCons reliy on automated metrics as proxy to human scores. To see how much are these methods limited by the proxy being imperfect, we replace the metrics with the human scores. This would make these methods illegible for real subset selection and thus should be treated rather as an oracle.

Item Response Theory. For the item response theory in the main paper, we use the product of difficulty b_x and discriminability a_x as item x's utility. Now, we include using the difficulty b_x, discriminability a_x, and feasability c_x alone. We also include Fisher information, which corresponds to the variance of the prediction (Equation 8) with respect to the model ability θ_m. The FI(x, θ) is defined as:

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log\frac{c_x}{1+\exp\left[-a_x(\theta-b_x)\right]}\Big|\theta\right] \quad (14)$$

$$= \mathbb{E}\left[\frac{a_x^2\cdot\exp\left(a_x\cdot(\theta_m+b_x)\right)}{(\exp(a_x b_x)+\exp(a_x\theta_m))^2}\Big|\theta\right] \quad (15)$$

This formula is a general form of that of Rodriguez et al. (2021); Lord and Novick (2008) which assumes $c_x = 1$ and $b_x = 0$. Because Fisher information is additive, we sum it across individual model abilities to obtain Fisher information content:

$$\text{FIC}(x) = \sum_{m\in\mathcal{M}} \text{FI}(x, \theta_m) \quad (16)$$

Diversity. Previously we have computed model output diversity using vector similarity based on multilingual embeddings of MiniLM-12-v2 (Reimers and Gurevych, 2019). However, this similarity can easily be replaced with simpler string-matching metrics, such as chrF (Popović, 2015), BLEU (Papineni et al., 2002), or Dice-Sørensen coefficient between unigrams. These metrics are averaged across pairwise comparisons between all model output pairs.

Sentinel (Perrella et al., 2024). The Sentinel metric has been proposed for the WMT Metrics Shared task to measure how much quality estimation relies on the source item difficulty rather than estimating the output quality. It is akin to MetricAvg$^{\text{src}}$ and also based on XLM-RoBERTa (Conneau et al., 2020), though without target-side aggregation (averaging).

Clustering. A common approach to benchmark creation is to maximize the diversity of the (source) items, such that the new test set covers diverse topics. This is not yet directly captured by any of the other methods. To directly optimize this, we embed each item into a vector, run k-means clustering with $k = B$, and for each each of the cluster choose the item to represent it whose embedding is closest to the cluster center. This ensures that items are chosen far away from each other in the vector space. The embeddings can be based either on the source texts or the model outputs.

Method	SPA	Method	SPA
Oracle subset (human, N = 10)	94.1%	k-means (sources)	92.3%
Oracle subset (human, N = 100)	95.3%	k-means (sources, weighted)	91.5%
Oracle subset (human, N = 1000)	96.1%	k-means (outputs)	91.8%
Oracle subset (MetricX, N = 10)	91.6%	k-means (outputs, weighted)	89.3%
Oracle subset (MetricX, N = 100)	91.7%	DiffUse	92.4%
Oracle subset (MetricX, N = 1000)	91.7%	Better than Random	91.2%
Greedy subset (human, N = 10, S = 10)	97.0%	Item Response Theory (difficulty)	89.0%
Greedy subset (human, N = 100, S = 10)	97.9%	Item Response Theory (discriminability)	88.7%
Greedy subset (MetricX, N = 10, S = 10)	92.2%	Item Response Theory (feasability)	84.7%
Greedy subset (MetricX, N = 100, S = 10)	91.8%	Item Response Theory (Fisher information)	92.5%
MetricAvg (human)	92.5%	COMET-instant-confidence (high error)	91.1%
MetricVar (human)	92.6%	COMET-instant-confidence (low error)	85.1%
MetricCons (human)	96.2%		
Diversity (Unigram)	92.4%	MetricAvg (MetricX) + MetricVar (MetricX)	93.1%
Diversity (ChrF)	92.1%	MetricAvg (MetricX) + MetricAvg (XCOMET)	92.6%
Diversity (BLEU)	92.9%	MetricVar (MetricX) + MetricVar (XCOMET)	93.5%
Sentinel-DA-src	92.0%	MetricCons (MetricX) + MetricVar (MetricX)	93.2%
Sentinel-MQM-src	92.2%	MetricCons (MetricX) + Diversity (LM)	93.1%

Table 8: Results for other methods described in Appendix A on WMT23 machine translation evaluation measured with soft pairwise accuracy. *MetricX* is MetricX-23 and *XCOMET* is XCOMET-XXL.

We extend this method further by making an assumption that examples close to each other in the vector space will lead to the same evaluation outcomes. Based on this, we pretend that all items in the cluster produce the same human scores as the selected item. In practice, this is done by assigning weight to each selected item based on the size of the cluster it represents.

DiffUse (Ashury Tahan et al., 2024). The DiffUse method is similar to clustering, but on the space of differences between model representations. For each item, we compute the distances between each pair. This then becomes a new vector with the size of $|\binom{\mathcal{M}}{2}|$ per each item. Then, these vectors are merged using hiearchical clustering to fit the predefined budget B. For the embeddings, we use the same multilingual encoder MiniLM-12-v2 (Reimers and Gurevych, 2019).

Mixture of Methods. Finally, we explore mixing methods together. Most of the methods score items independently to produce utilities. Our mixing approach is simple: For each considered utility function, rank the items based on the scores. Then, average the ranking with weights across multiple utility functions and select top-B.

Metric Uncertainty. Automated metrics, such as COMET (Rei et al., 2020), can be modified to not only provide the best estimate of the translation quality but also an estimate on its own error in the prediction. In the case of COMET-instant-confidence (Zouhar et al., 2025b), the additional prediction is expected absolute error from the true human score, which relates to epistemic uncertainty. Items with high uncertainty might be those difficulty to annotate and thus also lead to annotation noise, which we want to avoid. At the same time, the reason why these items are difficult to annotate for a metric might be because in these items the model outputs are most unclear or contested. In this method, we thus try to sample items with either high or low epistemic uncertainties of the automated metric.

A.2 Results

To test the additional methods, we again use the WMT23 dataset and average over languages and data proportions, such that the resulting numbers in Table 8 are comparable to that in the main paper.

As expected, using human scores when directly optimizing for the best subset leads to outstanding results. However, this is largely dependent on luck in the random subset generation and taking the maximum. Instead, building this subset greedily is more efficient and also holds more promise.

Unfortunately, replicating the same approach with automated metrics does not show promise above random. The best subset as meta-evaluated

with respect to an automated metric is not the best subset as meta-evaluated with respect to human scores. While MetricAvg and MetricVar do not benefit more from using human scores as opposed to the automated metric proxy (see Figure 4 for comparison), MetricCons performs much better than random and the rest of the previously discussed methods. This means, that further improvements in automated metrics will likely make MetricCons even more useful.

Using k-means clustering for subset selection is both a remedy to potential loss of diversity in the subset, and a well-performing method. The basic variant based on embedding just the sources and without any reweighting performs the best. This variant also works as a source-based selector and does not require the knowledge of model outputs, such as DiffUse with comparable performance. Similarly, even when allowing for immediate feedback with active sampling, the Better-than-random does not perform well in this setup.

Using different pairwise similarity metrics for Diversity does not improve over using embeddings. Selecting by epistemic uncertainty with COMET-instant-confidence also does not improve over other methods, though items with high uncertainty seem to be prefferable for evaluation to those with low uncertainty. The Sentinel metrics perform close to the output-based MetricAvg, which its imitates.

Lastly, combining multiple methods together (either same method with different metrics or two methods) using ranking averaging shows only diminishing improvements, which rarely surpasses the individual method's performance.

B Other Meta-evaluations

There are many ways to measure the usefulness of a subset for evaluation, depending on the goal, which we investigate in this section.

B.1 Meta-evaluations

The meta-evaluation measures how close some evaluation outcome of $\mathcal{Y} \subseteq \mathcal{X}$ is to that of $\mathcal{X}$.

(Soft) Pairwise Accuracy. In the main paper, we use soft pairwise accuracy, which combines the correctness of model ranking on $\mathcal{Y}$ with significance of that ranking. For each two models $m_1, m_2 \in \mathcal{M}$ we compute the p-value of the hypothesis that the averages scores of m_1 is higher than to that of m_2 (using a paired permutation test, N = 1000, Fisher, 1935; Good, 2000). These hypotheses are tested within $\mathcal{Y}$ and $\mathcal{X}$, denoted as $p^{\mathcal{Y}}_{m_1>m_2}$ and $p^{\mathcal{X}}_{m_1>m_2}$. This meta-evaluation captures the average similarity in confidences of pairwise comparisons:

$$\text{SPA}(\mathcal{Y}, \mathcal{X}) = \frac{1}{|\binom{\mathcal{M}}{2}|} \sum_{m_1,m_2 \in \binom{\mathcal{M}}{2}} 1 - |p^{\mathcal{X}}_{m_1>m_2} - p^{\mathcal{Y}}_{m_1>m_2}| \quad (17)$$

The soft pairwise accuracy has been proposed by Thompson et al. (2024) and simple pairwise accuracy is a special case where where the p-values are 0/1 based on if $m_1 > m_2$.

Correlations. Possibly the simplest evaluation is model ranking, which can be compared between the subset $\mathcal{Y}$ and the full set $\mathcal{X}$. This comparison can be done with a correlation, such as Pearson, Spearman, or Kendall variant b. The Pearson correlation stands out by taking the scale into consideration, which makes it sensitive to outliers.

Top-1 Match. In some cases, the purpose of the evaluation is to find the best model. For this, we measure how often the top model in $\mathcal{Y}$ to be the same as that of $\mathcal{X}$. Similar to Deutsch et al. (2022), this omits many of the evaluated examples to align the meta-evaluation with the practical goal of selecting the best model.

Model Average Error. In cases where we are not interested in the model ranking, but accurate absolute scores, we can measure the mean absolute error and (root) mean squared error between the model average on $\mathcal{Y}$ and the model average on $\mathcal{X}$.

Cluster Count. In the General WMT Shared Task (Kocmi et al., 2024a), we greedily compute clusters based on model ranking, as in Algorithm 1. Because all models in one cluster are statistically better than the models in the next cluster, the goal is to have evaluation that leads to as many cluster numbers as possible.

Top-1 Cluster Match. Evaluation with the top-1 match does not distinguishing between cases where there is only a small difference between the

Meta-evaluation	Random	MetricAvg	MetricVar	MetricCons	Diversity	DiffDisc
Soft pairwise accuracy	91.6%	92.6%	93.0%	**93.4%**	93.0%	92.6%
Pairwise accuracy	92.7%	93.4%	93.4%	**94.3%**	93.3%	93.2%
Pearson correlation	96.7%	96.5%	96.5%	97.5%	**97.5%**	96.9%
Spearman correlation	94.5%	95.0%	95.1%	**96.1%**	95.3%	95.1%
Kendall$_b$ correlation	85.3%	86.9%	86.9%	**88.6%**	86.7%	86.3%
Top-1 match	94.6%	95.0%	95.1%	**96.4%**	95.4%	96.2%
Cluster count	2.62	3.47	3.54	**3.68**	3.23	3.33
Top-1 cluster match	59.5%	**65.6%**	60.2%	64.2%	63.3%	65.1%
Mean average error ↓	**0.008**	0.055	0.046	0.031	0.036	0.028
Mean root squared error ↓	**0.010**	0.062	0.059	0.039	0.041	0.034

Table 9: Output-based subset selection methods on WMT23 meta-evaluated based on Appendix B.

Algorithm 1 Computation of number of clusters given an evaluated set of items.

Input: Models $\mathcal{M}$ **Output:** Number of clusters $|C|$

Add a system to the same cluster if it is not statistically distinguishable from the previous cluster.

$\mathcal{M} \leftarrow \textsc{Sort}(\mathcal{M}, \lambda m : -\textsc{Avg}(m))$
$C \leftarrow \langle \langle S_0 \rangle \rangle$
for $m \in \mathcal{M}_{>1}$
  if $\textsc{Wilcoxon}(C_{-1,-1}, m) < 0.05$
    $C.\textsc{append}(\langle m \rangle)$
  else
    $C_{-1}.\textsc{append}(m)$
return $|C|$

top models and cases where the top-1 model is the sole winner. To remedy this, we can compare the similarity between the top-1 clusters (computed as in Algorithm 1) based on $\mathcal{Y}$ and $\mathcal{X}$. For this, we use the Sørensen–Dice coefficient: $\frac{2\times|C_1^{\mathcal{Y}} \cap C_1^{\mathcal{X}}|}{|C_1^{\mathcal{Y}}|+|C_1^{\mathcal{X}}|}$.

Active sampling. Ruan et al. (2024) propose a constrained active sampling framework that iteratively selects examples for inclusion in $\mathcal{Y}$. This method combines a learned model for quality prediction, systematic sampling to ensure diversity across quality levels, and a constrained controller to reduce redundancy. However, due to its active sampling nature, it requires immediate feedback of the human score upon selecting an item, which is not compatible with the selection of the whole subset at once prior to human evaluation.

B.2 Results

The results are shown in Table 9 for output-based subset selection methods of the main paper.

The biggest gains over the random selection can be seen for the cluster count, with up to +1 cluster when using MetricCons. This improvement is meaningful, because in the context of a shared task, having an extra significance clusters allows for more statistically justified claims. Further, even though the improvements on correlation-based meta-evaluations over random are not by a large margin, they are consistent. The only exceptions are the differences from the model averages. However, this is not unexpected and rarely the goal of the meta-evaluation. Central limit theorem shows that the mean of observed independent random variables (item scores) converges to the true mean (model score average) at least as fast as $x^{-1/2}$ converges to 0. Sampling not randomly, such as by difficulty (MetricAvg), creates a bias in the average of the random variable, which results in higher error from the mean on the whole set.

C Implementation Details

Item Response Theory. The item response theory models usually take the form of a logistic regression and are optimized with stochastic variational inference (Wu et al., 2020; Rodriguez et al., 2021), which explicitly take into account the distributional priors and dependencies between the latent variables. The priors for item difficulty (b), item discriminability (a), and model ability (θ) is the normal distribution and we implement the model in `py-irt` (Lalor and Rodriguez, 2023).

PreCOMET. For item utility distillation we use PreCOMET, which is losely based on COMET (Rei et al., 2020). This model starts as XLM-RoBERTa (Conneau et al., 2020), with

attached regressor head (multi-layer perceptron, $768 \times 2048 \times 1024 \times 1$), and optimized with Adam ($\text{lr} = 1.5 \times 10^{-5}$) with weight decay (0.95). The model is trained for 5 epochs with the effective batch size of 128 on training data of WMT before 2023 (50k sources). The item utilities are computed with respect to human scores and not automated metrics, since during training human scores are available.

Output Diversity. We compute the diversity in model outputs using inner products of embeddings in Equation (7) based on multilingual MiniLM-12-v2 (Reimers and Gurevych, 2019).

	Dataset	#Models	#Items	Random	MetricAvg	MetricVar	MetricCons	Diversity	DiffDisc
WMT24	Cs→Uk	11	1954	92.5%	93.7%	**96.0%**	92.1%	93.3%	93.9%
	En→Cs	15	571	90.3%	86.5%	87.0%	89.6%	**92.1%**	89.0%
	En→Es	13	634	90.7%	91.0%	90.5%	90.5%	90.1%	**91.2%**
	En→Hi	10	634	89.6%	86.1%	87.8%	**92.5%**	90.3%	91.7%
	En→Is	10	634	95.8%	96.1%	95.3%	92.2%	**96.4%**	94.3%
	En→Ja	12	634	85.1%	82.0%	**85.5%**	84.5%	85.3%	84.7%
	En→Ru	13	634	91.1%	90.4%	91.1%	92.7%	**92.8%**	92.1%
	En→Uk	10	634	90.7%	**93.8%**	88.0%	92.4%	93.4%	91.6%
	En→Zh	12	634	**87.9%**	87.3%	86.7%	86.8%	84.3%	86.7%
	Ja→Zh	14	559	93.0%	95.6%	**96.4%**	94.0%	94.3%	94.0%
	Cs→Uk	13	1009	90.0%	89.1%	90.4%	**94.4%**	92.5%	91.6%
WMT23	De→En	13	509	91.0%	91.5%	91.4%	92.7%	**93.9%**	92.6%
	En→Cs	15	1098	87.7%	87.4%	90.1%	**91.8%**	90.1%	89.5%
	En→De	12	549	92.1%	91.8%	92.1%	91.9%	91.7%	**92.3%**
	En→Ja	16	1098	93.3%	**95.2%**	94.6%	94.5%	94.4%	94.4%
	En→Zh	15	1098	92.6%	94.5%	93.8%	94.7%	**94.9%**	94.8%
	He→En	12	820	92.9%	94.9%	**96.4%**	95.6%	92.8%	94.1%
	Ja→En	17	1120	92.9%	**95.8%**	95.1%	92.2%	95.2%	94.2%
	Zh→En	15	884	91.4%	93.3%	**94.4%**	90.3%	91.1%	91.0%
	Cs→En	11	561	**84.7%**	73.9%	78.4%	83.0%	77.9%	84.6%
WMT22	Cs→Uk	10	819	90.6%	87.4%	87.4%	91.5%	88.5%	**93.1%**
	De→En	9	601	**80.2%**	76.7%	76.4%	74.9%	78.5%	75.6%
	En→Cs	10	1205	89.8%	90.1%	**95.7%**	92.1%	89.9%	90.1%
	En→De	14	1315	91.8%	**94.5%**	**94.5%**	90.5%	92.8%	92.4%
	En→Hr	8	1107	92.6%	94.2%	95.2%	**96.8%**	92.6%	94.0%
	En→Ja	13	1181	85.9%	83.2%	87.2%	86.9%	85.0%	**87.3%**
	En→Ru	15	1315	95.0%	95.9%	95.1%	95.2%	94.8%	**96.3%**
	En→Uk	8	1209	91.5%	92.1%	**95.3%**	94.9%	90.4%	93.1%
	En→Zh	12	1181	84.1%	83.1%	**89.8%**	80.4%	81.0%	88.9%
	Ru→En	10	1019	85.6%	86.5%	86.7%	83.4%	82.1%	**87.3%**
	Sah→Ru	2	1023	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	Uk→En	9	856	86.7%	87.2%	87.8%	**89.2%**	82.4%	84.7%
	Zh→En	14	1875	90.2%	92.2%	**95.2%**	93.1%	93.9%	91.4%
	En→De	13	529	83.8%	87.7%	85.7%	**90.0%**	85.0%	85.4%
WMT21	En→Ru	14	512	88.3%	**94.8%**	93.0%	89.3%	92.7%	92.9%
	Zh→En	13	529	83.3%	83.4%	77.6%	80.3%	82.3%	**84.9%**
	De→En	19	653	85.9%	85.2%	87.2%	**89.4%**	84.7%	84.7%
	En→Cs	10	988	97.4%	96.4%	98.2%	**98.7%**	97.4%	98.5%
	En→De	13	527	86.8%	91.1%	89.5%	**93.5%**	90.1%	91.0%
	En→Is	11	838	96.0%	92.4%	**96.9%**	96.3%	96.4%	96.2%
	En→Ja	15	878	94.0%	89.1%	94.2%	**95.2%**	88.3%	93.3%
	En→Ru	14	527	86.7%	**93.7%**	91.6%	91.4%	87.9%	91.2%
	Zh→En	13	650	81.3%	87.3%	87.7%	83.0%	**89.4%**	83.3%
WMT20	Zh→En	8	2000	88.4%	**94.1%**	91.8%	86.4%	88.1%	91.3%
	En→De	7	1418	90.3%	86.3%	85.1%	91.0%	85.0%	**92.2%**
WMT19	Kk→En	11	1000	89.5%	86.0%	**92.8%**	92.1%	84.3%	92.7%
	De→En	16	1948	89.0%	88.8%	**92.9%**	90.2%	89.2%	90.3%
	Gu→En	11	1016	92.2%	85.0%	92.2%	**95.4%**	89.6%	93.0%
	Lt→En	11	1000	92.6%	91.8%	91.6%	89.6%	92.0%	**93.7%**
SummEval	Relevance	11	100	94.0%	95.7%	**95.9%**	95.5%	94.2%	93.3%
	Coherence	11	100	94.7%	**95.6%**	94.5%	95.0%	95.3%	95.1%
	Consistency	11	100	90.2%	91.8%	91.2%	**92.5%**	92.0%	91.0%
	Fluency	11	100	90.2%	93.9%	92.0%	**94.1%**	91.1%	93.1%
	Sum	11	100	95.5%	**96.4%**	95.8%	96.2%	96.0%	94.5%
	Mul	11	100	95.4%	**96.1%**	94.8%	**96.1%**	95.4%	93.8%

Table 10: Individual subset selection results for machine translation evaluation (WMT) and summarization evaluation (SummEval) measured with soft pairwise accuracy averaged over data proportions from 5% to 50% (WMT) and 25% to 75% (SummEval). **Bold** numbers indicate best in evaluation category (correlation or clusters) within the row. Random is average over 100 runs.