Contextual Metric Meta-Evaluation by Measuring Local Metric Accuracy

Anonymous ACL submission

Abstract

Meta-evaluation of automatic evaluation metrics—assessing evaluation metrics themselves—is crucial for accurately benchmarking natural language processing systems and has implications for scientific inquiry, production model development, and policy enforcement. While existing approaches to metric meta-evaluation focus on general statements about the absolute and relative quality of metrics across arbitrary system outputs, in practice, metrics are applied in highly contextual settings, often measuring the performance for a highly constrained set of system outputs. For example, we may only be interested in evaluating a specific model or class of models. We introduce a method for contextual metric meta-evaluation by comparing the *local metric accuracy* of evaluation metrics. Across translation, speech recognition, and ranking tasks, we demonstrate that the local metric accuracies vary both in absolute value and relative effectiveness as we shift across evaluation contexts.

1 Introduction

Meta-evaluation of automatic evaluation metrics—assessing evaluation metrics themselves—is crucial for accurately benchmarking natural language processing systems (Zhou et al., 2022). Because metrics are central to scientific inquiry, production model development, and policy enforcement (Kocmi et al., 2021), there is a constant need for new approaches to evaluating system outputs (Novikova et al., 2017).

Although current methods for metric meta-evaluation commonly take a *global* perspective, reporting the performance of a metric across arbitrary system outputs, coming from any system (Stanojević et al., 2015; Przybocki et al., 2009), practical metric meta-evaluation is highly contextual, measuring the performance for a highly constrained set of system outputs. For example, we may only be interested in evaluating a specific model or class of models. From a model development perspective, we may be interested in a metric that is sensitive to model outputs coming from partially trained models at the beginning of the development cycle (when the outputs are far from the target distribution or close to random); such a metric may struggle to differentiate between outputs from fully trained or more effective models. This is highly reflective of the results found by Fomicheva and Specia (2019), who show that metric performance varies significantly across different levels of translation quality. Thus, using the same metric throughout the development process may lead to biased or incomplete evaluations and possibly pruning earlier models, which may have a better performance when fully trained.

	context			
metric	**X**	**Y**	**Z**	**global**
A	0.9	0.9	0.3	0.7
B	0.7	0.7	0.7	0.7
C	0.3	0.3	0.9	0.5

Table 1: Contextual metric meta-evaluation. When comparing metrics A, B, and C, traditional meta-evaluation focuses on global accuracy across arbitrary inputs. Local metric accuracy can vary by evaluation contexts X, Y, and Z.

To illustrate the difference between global and contextual metric meta-evaluation, we constructed a toy meta-evaluation for three metrics across three contexts (Table 1). The values in the table represent the accuracy of three metrics (A, B, and C) under three different contexts (X, Y, and Z) as well as the global accuracy across the different contexts. By looking at the average, we might think that A and B are equally accurate. However, when inspecting accuracy within individual contexts, we can see that selecting the most appropriate metric is far less straightforward. For example, if we want a metric that best generalizes across different contexts, we want to pick B over A even though their global accuracies are equal. However, if we want to specifically measure outputs in context Z, then we would want to pick C as it is especially sensitive to system outputs in that context, despite it having the lowest global accuracy. This suggests that comparing a metric on its global accuracy may deviate from local accuracy, which may be more relevant in a contextual setting.

To improve contextual metric meta-evaluation, we propose analyzing metrics across different evaluation contexts and measuring their *local metric accuracies*. By evaluating metrics across three different machine learning tasks—machine translation, automated speech recognition, ranking—we show that the metric accuracy, which measures the ability of a metric to accurately assign the true preference between a pair of system decisions, changes as the context of the output space changes. We also show that the metric accuracy changes both in absolute value and relative ordering across the different contexts. In contrast with existing work on metric meta-evaluation relies heavily on costly and time-consuming explicit human feedback (Fabbri et al., 2021; Liu et al., 2016), our method uses output perturbations (Sai et al., 2021; He et al., 2023) to obtain the true ordering between a pair of system outputs without the need of human supervision. Overall, we show that measuring local metric accuracies is a straightforward methodology to provide a more contextual understanding of evaluation metrics which complements existing global metric meta-evaluation methods.

2 Related work

Our work connects to the broader literature on meta-evaluation, which has been approached in various ways, highlighting the complexity and the necessity of this task. For example, the Workshop of Statistical Machine Translation (WMT) has focused on evaluating the utility of metrics in machine translation since 2008, where participants submit automated metrics for validation against human feedback (Callison-Burch et al., 2008). However, human feedback can be subjective and susceptible to social biases (Sun et al., 2022). Noting these limitations, Xiao et al. (2023) propose a theory-driven meta-evaluation framework rooted in measurement theory for NLG metrics. Their work highlights issues in human evaluation including a lack of validation, standardization, and consistency.

Our use of output perturbation is inspired by prior work in testing metric robustness. Chen and Eger (2023) proposed a preference-based adversarial attack framework using targeted perturbations to evaluate the robustness of NLI-based and BERT-based metrics, finding that NLI-based metrics are more robust in summarization but not in machine translation. Sai et al. (2021) extends perturbation-based robustness testing by creating templates targeting specific criteria such as jumbled word order to test fluency. Chen et al. (2019) assessed QA metrics by converting multiple-choice datasets into free-response formats, highlighting the need for BERT-based metrics. Additionally, Valcarce et al. (2018) evaluated the robustness of ranking metrics against incompleteness by introducing sparsity to system outputs. Our paper adopts similar perturbation techniques to assess the preference-based evaluation capability of different metrics, eliminating the need for costly and time-intensive explicit human feedback.

Although metric meta-evaluation is often done on a global level, previous work indicates that the reliability of a metric changes from the system-level to the decision-level (Reiter and Belz, 2009; Stent et al., 2005). Though some research has investigated metric performance for different contexts based on output sources (i.e., models) or output qualities (Mathur et al., 2020; Novikova et al., 2017), our work addresses the lack of a systematic review of contextual meta-evaluation and how to conduct it.

3 Local accuracy

To formalize local metric accuracy, we introduce the following notation. Let $\mathcal{X}$ be the set of all possible system inputs (e.g., for MT, all possible strings from the source language) and $\mathcal{Y}$ the set of all possible system outputs (e.g., for MT, all possible strings from the target language). We define $X \subset \mathcal{X}$ to be the subset of system inputs observed in a specific context (e.g., for MT, a sample of source sentences from a specific university). Similarly, $Y_x \subset \mathcal{Y}$ is the subset of system decisions for $x \in X$ observed for X in a specific context (e.g., for MT, a set of translations generated by a set of candidate systems). In addition, we have access to a perturbation function that, with high probability, degrades the utility of a decision y (e.g., dropping a random word from a translated input). Let Q_x

be the set of pairs decisions conditioned on an input x and their corresponding degraded version: $Q_x = \{\langle y, y' \rangle\}_{y \in Y_x}$.

An evaluation metric $\mu : \mathcal{X} \times \mathcal{Y} \to \Re$ generates a scalar number reflecting the performance according to some system property that we want to measure (e.g. correctness of translation). Each metric is an approximation of μ^*, the ideal evaluation metric (i.e., the true utility of an output). Given any two pairs of system outputs, $\mu*$ will always be able to determine the true ordering of the two outputs. In cases where we intentionally perturb y to obtain y', we know that $\mu^*(x, y) > \mu^*(x, y')$. Under the assumption that μ approximates μ^*, we want to compute how often $\mu(x, y) > \mu(x, y')$. As suggested by Kocmi et al. (2021), we focus on the ability of μ to reproduce the ordering of decisions, rather than the magnitude of the difference between $\mu(x, y)$ and $\mu(x, y')$. From this, we define the pointwise local metric accuracy, conditioned on an input x as,

$$\textsc{Acc}_\mu(Q_x) = \frac{1}{|Q_x|} \sum_{\langle y, y' \rangle \in Q_x} \mathbb{1}\left[\mu(x, y) > \mu(x, y')\right] \tag{1}$$

This measures the ability of a metric to reproduce the true ordering of perturbations for a specific input x. We define the local metric accuracy across all contexts as,

$$\textsc{Acc}_\mu(Q) = \frac{1}{|X|} \sum_{x \in X} \textsc{Acc}_\mu(Q_x) \tag{2}$$

where $Q = \cup_{x \in X} Q_x$. This measures the local metric accuracy across a sample of system inputs, as we may have in a standard evaluation set.

We are interested in testing two hypotheses with respect to local metric accuracy.

H1: The absolute local metric accuracy, $\textsc{Acc}_\mu(Q)$, of a metric μ changes as the context changes.

Evidence supporting this hypothesis suggests that existing evaluation methods focusing on global metric accuracy obscures how metric accuracy varies across different contexts.

H2: The ordering of a set of metrics by local metric accuracy changes as context changes.

In other words, the total ordering of all metrics by local metric accuracy within a context changes as the context Q changes. Evidence supporting this hypothesis suggests that choosing an appropriate metric to benchmark compare system outputs largely depends on the context.

4 Methods and Materials

4.1 Tasks, dataset, and metrics

We performed our evaluation on three different tasks: Machine Translation (MT), Automated Speech Recognition (ASR), and Ranking. Table 2 details the dataset and metrics that we used in our experiments. For each task, we used readily available system outputs to improve reproducibility. For each metric, we employed their respective official implementations or, when unavailable, the most widely used implementation with default parameters. For any neural metric computation, we used a NVIDIA RTX A6000 GPU. For BLEU, we used `nltk`'s `sentencebleu` implementation. We also used `nltk`'s implementation for METEOR. For ROUGE[1] and BERTSCORE[2], we have used the implementation released by their respective authors. For BLEURT[3], COMET[4], CHRF[5] and UNITE[6], we have used their official implementations via the `evaluate` library on HuggingFace. We used the `jiwer`[7] Python package to compute the ASR metrics. We used the `trec_eval`[8] to calculate the ranking metrics.

We adopt the category with the highest number of contexts for each task. The abundance of contexts allowed us to identify trends in metric behavior across a broader range of items and helped us identify supporting evidence for or against our hypotheses.

4.2 Perturbation techniques

To test our hypotheses, we applied a perturbation function that degrades the utility of a system output y and its corresponding degraded version y'. Thus, we know that the quality of y' under a specific task is worse than y with a high probability. For the system outputs belonging to the machine translation and automated speech recognition tasks, we perturbed y by removing 20% of the words in the outputs, rounded to the nearest integer.

Our perturbation technique is a simplification of He et al. (2023), who synthesize a range of perturbations that closely mimic human or machine

[1] https://github.com/google-research/google-research/tree/master/rouge

[2] https://github.com/Tiiiger/bert$_s core$

[3] https://github.com/google-research/bleurtreadme

[4] https://unbabel.github.io/COMET/html/models.html

[5] https://github.com/mjpost/sacreBLEUchrf–chrf

[6] https://huggingface.co/Unbabel/unite-mup

[7] https://github.com/jitsi/jiwer

[8] https://github.com/usnistgov/trec_eval

Task	Dataset	Metrics
MT	Over 150,000 system outputs and reference translation from 62 different MT systems submitted to the WMT metrics task from year 2023 (Freitag et al., 2023) for the source-target language pairs English-Russian (en-ru), English-German (en-de), Chinese-English (zh-en). The subsets that are available are YEAR, DOMAIN, and SYSTEM.	BLEU, ROUGE-1, ROUGE-2, ROUGE-L, METEOR, BERTSCOREP, BERTSCORER, BERTSCOREF1, COMET, BLEURT, CHRF, UNITESRC, UNITEREF, UNITEUNIFIED
ASR	Over 33,000 system outputs from six different ASR models on ESPnet (Watanabe et al., 2018) on the LibriSpeech 100 dataset (Panayotov et al., 2015). The subsets that are available are SYSTEM, SPEAKER ID, GENDER, and QUALITY.	Word Error Rate (WER), Match Error Rate (MER), Word Information Lost (WIL), Word Information Preserved (WIP), Character Error Rate (CER)
Ranking	Ranked list of top-100 items retrieved by 21 recommender algorithms provided by Valcarce et al. (2018) on the MovieLens1M dataset (Harper and Konstan, 2015) submitted to TREC (Buckley and Voorhees, 2004). We were able to segment the outputs by ALGORITHM.	Mean Average Precision (MAP), Precision@R, where R is the number of relevant documents (RPREC), Reciprocal Rank (RECIP_RANK), Interpolated Precision at Recall Level X (for $X = \{0.0, 0.1, 0.2, 0.3, 0.4\}$) (IPREC_AT_RECALL_X), Precision@K (P_K), Recall@K (RECALL_K), nDCG@K (NDCG_CUT_K) (where $K = 5, 10, 15, 20, 30$)

Table 2: Datasets and metrics used for different tasks

errors. However, we want added simplicity and generalizability to languages other than English, so we refrained from doing perturbations that are semantically informed, such as removing articles and prepositions, verb lemmatization, or negation.

For the system outputs belonging to the ranking task, we perturbed y by shuffling the rankings within the top-100 items for each user-system pair.

4.3 Hypothesis Testing

In order to test H1, we plotted the metric accuracies $\text{ACC}_\mu(Q)$ for each task across different contexts within the selected context category Q as a line graph, such that we can visualize how the metric's capability of differentiating between y and y' changes as the context changes by observing the slopes and overlaps between the lines. To further investigate the association between the context Q and the metric accuracy $\text{ACC}_\mu(Q)$, we used the χ^2 test of independence of variables (Pearson, 1900) in a contingency table (Pearson, 1904). We will compare the resulting p-values to the significance level of $\alpha = 0.05$ to understand whether the changes in metric accuracy $\text{ACC}_\mu(Q)$ across the different contexts Q are statistically significant.

To test H2, we computed the Kendall's τ (Kendall, 1938) between a two rankings of metrics according to local metric accuracy under two contexts. This helps us quantify how the total ordering of the metrics changes as the context changes. In order to emphasize the metric selection task, we adopt version of Kendall's τ that weighs changes at the higher in the ranking more than those lower in the ranking (Shieh, 1998). Specifically, we use a hyperbolic weighing that maps each rank r to weight $\frac{1}{r+1}$.

5 Results

5.1 Machine Translation

We visualize the metric accuracies for the machine translation metrics under the different SYSTEM contexts, as shown in Figure 1a. We observe that the line for each metric changes as we change the context, as indicated by the varying slopes of the lines. The results of the χ^2 test indicate that the difference in the metric accuracies across the different context is statistically significant, supporting H1 for MT.

Figure 1a also contains intersections between lines corresponding to different metrics, indicating that there is a change in the relative position of each metric in the different contexts and hence signifying a change in the total ordering of metrics by local accuracy across contexts, supporting H2. This is further strengthened in Figure 1b, where the τ values show that the correspondence between the pairs of metric accuracy rankings varies considerably for each pair of SYSTEMs. If the τ values were consistently close to 1, there would not be support for H2. Instead, we find that τ values cluster according to similarity of context.

5.2 Automated Speech Recognition

For ASR, we report the local metric accuracy under the different SPEAKER IDs which come from different dataset QUALITY contexts (CLEAN/RAW). We plot the local metric accuracy for contexts associated with different contexts shown in Figure 2a. We observe that the lines corresponding to each metric are not straight, which indicates that the absolute local accuracy for each metric changes with context, supporting H1. Our χ^2 test results confirm that the difference in the local metric accuracies across the different contexts is statistically significant, providing evidence supporting H1 for ASR.

Interestingly, we do not observe the same overlap between the lines corresponding to the different metrics as we did for MT. This, along with the consistently high values of τ in Figure 2b, indicates that there is not evidence supporting H2 for ASR.

5.3 Ranking

Plotting the metric accuracies for the ranking metrics for the different ALGORITHMs in Figure 3a, we can first observe that none of the lines corresponding to the different metrics are straight lines, which supports H1. The large fluctuations within each line suggest that the changes in the absolute local accuracies for each metric are rather significant. The χ^2 test results shows a statistically significant change, providing evidence in support of H1.

Furthermore, we can see that overlaps exist between the different lines corresponding to the different metrics, similar to the observation we made in the MT case. This indicates that the total ordering of metric accuracies changes as the context changes, supporting H2. The τ results (Figure 3b) show clustering by algorithm, as with MT.

6 Discussion

6.1 H1: Absolute Local Accuracies

The results in Section 5 generally provide evidence supporting H1, as our experiments consistently show that the local metric accuracy changes as the context changes.

We can observe that the local metric accuracy for a context is related to the average quality of outputs in that context. For example, in the ranking setting, the most effective system according to MAP is SLIM, which is also the context whose perturbed outputs are easiest to distinguish. Conversely, perturbed outputs in the random ranker are more difficult for all metrics to distinguish. This is because our perturbation method catastrophically degrades good outputs and bad outputs are already poor and difficult to make demonstrably worse. We will return to this in Section 6.3.

These results suggest that evaluators may be interested in the stability of local metric accuracy when selecting a metric. A metric that is more stable with respect to local metric accuracy is more predictable when deployed under a new context and the probability of selecting the wrong system is consistent. This is especially important if we consider a new evaluation context where poor local metric accuracy puts users—or a vulernable subsgroup of users—at risk. Work in robust machine learning provides existing methods for designing metrics stable across context changes (Yuan et al., 2024).

In addition to stability, we can organize metrics according to systematic behavior in local metric accuracy. For example, in Figure 1a, more complex embedding-based and model-based evaluation metrics generally perform better than the simpler lexical-based metrics (Zhang et al., 2019; Freitag et al., 2022). More complex metrics cluster on the top of the figure while simpler metrics occupy the bottom regions; any overlap occurs mostly within a specific category. Such analysis allows evaluators to understand the empirical relationships between metric ensembles. Although picking the best metric might involve selecting a metric occupying the top of the figure, there may be contexts in which local metric accuracies are close enough to allow flexibility in selecting metrics with lower local metric accuracy.

More generally, we can consider multi-objective metric development. For example, since embedding- and model-based methods are more time-intensive and computationally costly compared to the lexical-based methods, adopting simpler and cheaper metrics when local metric accuracies are comparable (e.g., early in model development) would result in cost savings and faster iteration. Beyond cost and local metric accuracy, one can imagine local versions of metric interpretability, metric engineering overhead, metric optimizability, and other criteria when conducting for contextual meta-evaluation.

6.2 H2: Relative Local Accuracies

Although the observations in Section 5 errs toward accepting H2, the evidence from our experi-

(a) Local metric accuracy across the different contexts

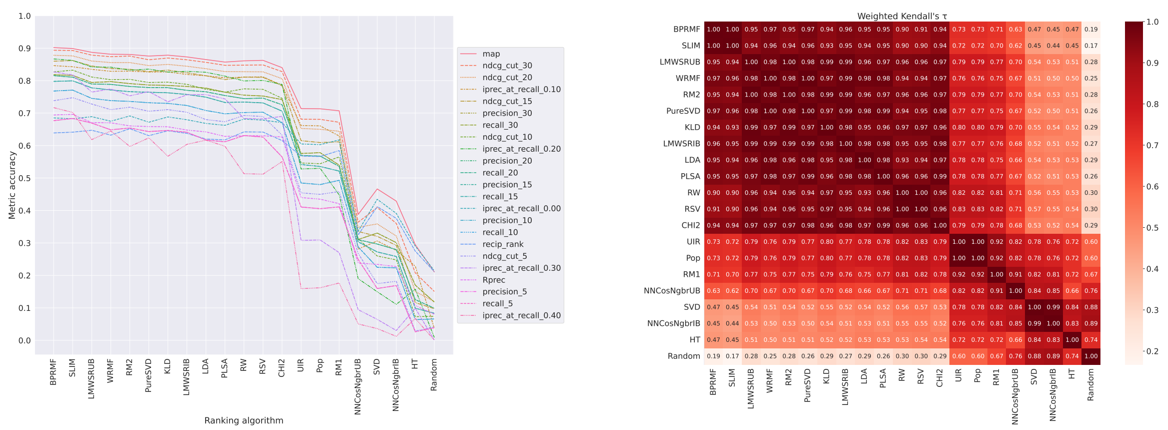
(b) Weighted Kendall's τ of metrics ordered by local accuracy between different contexts

Figure 1: Machine Translation. Contexts provided by translation system.

(a) Local metric accuracy across the different contexts

(b) Weighted Kendall's τ of metrics ordered by local accuracy between different contexts

Figure 2: Automatic Speech Recognition. Local metric accuracy across different Speaker IDs. (a) Speaker IDs to the left of the grey line come from the QUALITY=CLEAN LibriSpeech-100 dataset, while the Speaker IDs to the right of the grey line come from the QUALITY=OTHER LibriSpeech-100 dataset.



(a) Local metric accuracy across the different contexts

(b) Weighted Kendall's τ of metrics ordered by local accuracy between different contexts

Figure 3: Ranking. Metric accuracy for Ranking metrics across the different systems.

ments on the ASR task in Section 5.2 suggest that it strongly depends on the nature of the task and the metrics. Evaluating ASR is relatively straightforward, where the ambiguity of correct answers is low, unlike MT or ranking where two outputs (e.g., translations or permutations) can be equally good (Wieting et al., 2019). Hence, the metrics are commonly used to benchmark ASR systems only slightly vary in the construct they are trying to measure, and they are all operationalized following similar statistical methods.

Figures 1b and 3b indicate that there are groups of contexts where the relative reliability of metrics is similar. When contexts can be structured according to metric accuracy, ones can adopt a fixed evaluation metric. This has practical implications in terms of engineering and development overhead or, in the case of model-based metrics, model development cost. *Predicting* the similarity in local metric accuracy ordering (i.e., the cells in Figures 1b and 3b) is an important task because it allows evaluators to confidently adopt an evaluation metric without conducting contextual meta-evaluation. Predictive features include any metadata we have about the contexts. For example, in ranking, Valcarce et al. (2018) categorize ALGORITHMs into different families of techniques: matrix factorization (SVD, PURESVD, BPRMF, WRMF), neighborhood-based (CHI2, KLD, RSV, ROCCHIO'S WEIGHTS).

6.3 Methodology

Although our results demonstrate that local metric accuracy analysis can provide insight into metric behavior, there are several opportunities for improving the methodology. First, our perturbations, while reliable in generating output degradations, may result in outputs that are easily detected by metrics, especially for highly effective systems. Moreover, perturbed outputs may be sufficiently different as to be unlikely to occur in a specific context. For example, if we are evaluating in the context of highly effective MT systems, a translation with a missing word is very unlikely by any highly effective MT system, even though we know it is lower quality. In order to address this, developing perturbation methods that reliably degrade performance *and* are likely to occur within a context will be important for future local metric accuracy development. This is related to synthesizing hard negative examples in the contrastive learning literature (Kalantidis et al., 2020). Alternatively, we can consider non-perturbation data, perhaps from human annotators, although this compromises the cost-effectiveness of output perturbation.

In order to help with clarity, we focused on contexts that were interpretable, which contexts are relevant depends on the broader model evaluation environment. Focusing on models, as we did for MT and ranking, emphasizes contexts that reflect iterative model development and refinement within a narrow set of constraints (i.e., the particular model being evaluated). If we are benchmarking a diverse set of systems, we are interested in comparing a broader set of possible outputs than those from a single system. In cases where we are designing a metric agnostic to a particular context, we may be interested in robust performance across arbitrary contexts. While this is similar to global analysis, a more rigorous and formal approach to context selection, such as found in the distributionally robust machine learning literature (Duchi et al., 2018), may be more appropriate.

7 Conclusion

We introduce the notion of local metric accuracy and demonstrate how to use it to conduct contextual metric meta-evaluation. Our results show that both the absolute and relative local accuracy of a metric varies as we vary context, though this depends on the nature of the task. Based on our results, we believe that by moving beyond global metric meta-evaluation, we can achieve a more accurate understanding of metric performance, which in turn increases the reliability of the evaluations and provide actionable insights for improving NLP systems.

8 Limitations

As mentioned in Section 4.2, our experiments adopted relatively simple perturbation methods in order to cover a wide range of tasks and guarantee degradation. In future work, we plan to explore more more task- and language-specific methods developed in the NLP community (Sai et al., 2021; Chen and Eger, 2023; He et al., 2023).

We also compute local accuracy by uniformly weighting all output-perturbation pairs. In reality, different outputs have different probabilities of occurring in a specific context. These probabilities should be incorporated into the accuracy calculation to provide a more reliable estimate of local metric accuracy. Estimating the distribution over outputs for a specific context itself is a difficult research question which we plan on addressing in future work.

References

Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron Shaw Fordyce. 2008. Proceedings of the third workshop on statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

J. C. Duchi, P. W. Glynn, and H. Namkoong. 2018. Statistics of robust optimization: A generalized empirical likelihood approach. *CoRR*, abs/1610.03425.

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.

Marina Fomicheva and Lucia Specia. 2019. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.

Markus Freitag, Nitika Mathur, Chi kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Dan 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 576–626, Singapore.

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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21798–21809. Curran Associates, Inc.

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

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

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Karl Pearson. 1904. *On the theory of contingency and its relation to association and normal correlation*. Drapers' Company research memoirs. Cambridge University Press.

Mark Przybocki, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. 2009. The nist 2008 metrics for machine translation challenge—overview, methodology, metrics, and results. *Machine Translation*, 23:71–103.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Grace S Shieh. 1998. A weighted kendall's tau statistic. *Statistics & probability letters*, 39(1):17–24.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International conference on intelligent text processing and computational linguistics*, pages 341–351. Springer.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2018. On the robustness and discriminative power of information retrieval metrics for top-n recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pages 260–268.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2024. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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, Seattle, United States. Association for Computational Linguistics.