Analyzing and Evaluating Correlation Measures in NLG Meta-Evaluation

Anonymous ACL submission

Abstract

The correlation between NLG automatic evaluation metrics and human evaluation is the most critical criterion for assessing the capability of an evaluation metric. However, different grouping methods and choices of correlation coefficients result in at least 12 types of correlation measures. For a long time, little has been known about their characteristics. Therefore, this paper illustrates the relationships between different correlation measures and demonstrates how the degree of data discretization affects their values through statistical simulations. Additionally, we designed algorithms to evaluate the discriminative power and ranking consistency of 12 correlation measures using empirical data from 6 datasets and 32 evaluation metrics, uncovering many interesting conclusions.

1 Introduction

Automatic evaluation metrics (e.g. BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020)) are widely used in Natural Language Generation (NLG) evaluation, and the evaluation of these evaluation metrics is known as NLG meta-evaluation. In NLG meta-evaluation, human evaluation is generally considered the gold standard. Therefore, the correlation with human evaluation is the most crucial criterion for assessing the performance of an evaluation metric. However, due to the existence of different grouping methods (e.g., system level (Bhandari et al., 2020), dataset level (Fu et al., 2023)) and different correlation coefficient functions (e.g., Pearson's r, Spearman's ρ), the measurement of correlation is not uniform. Due to a lack of understanding of the characteristics of different correlation measures, researchers often simply follow the practices of past work or authoritative competitions such as WMT in practice. However, many papers do not even clearly describe the correlation measures used, such as not reporting the grouping method, not to mention explaining why these measures are selected. Furthermore, the correlation measures used in authoritative competitions frequently change: WMT22 (Freitag et al., 2022) used segment-level correlations with three different grouping methods, while WMT23 (Freitag et al., 2023) used only one. This makes researchers confused about how to select correlation measures.

On the other hand, as Large Language Models (LLMs) are increasingly used for automatic evaluation, the selection of correlation measures has become more important and complex. This is crucial because numerous LLM evaluators have been proposed, including both prompting proprietary LLMs for NLG evaluation (Liu et al., 2023; Chiang and Lee, 2023; Kocmi and Federmann, 2023) and fine-tuned LLM evaluators (Wang et al., 2023; Xu et al., 2023; Jiang et al., 2023; Li et al., 2023). The current confusion surrounding correlation measures can severely hinder performance comparisons. Additionally, unlike traditional continuous evaluation metrics, LLM evaluators can score on a given scale according to user needs (e.g., 1-5, 0-100), which allows their output scores to contain more ties. This fact affects the fairness of comparisons under certain correlation measures (Deutsch et al., 2023).

It is by no means easy to strictly determine whether a particular correlation measure is reasonable, as this depends on the specific scenario and the researcher's preferences. For example, Deutsch et al. (2023) believes that in the evaluation of machine translation using MQM, the fine-grained human evaluation ties should be trusted and used to reflect whether a metric can correctly evaluate ties, while existing various Kendall correlation coefficients cannot handle this preference. However, in coarse-grained human evaluation (such as Likert-scale rating), human evaluation ties may not be trustworthy. Therefore, this paper does not directly discuss whether a particular correlation measure is reasonable but instead analyzes and presents their characteristics to enhance understanding. Our study revolves around the following research questions:

- **RQ1** (§4.3): What is the relationship between different correlation measures?
- **RQ2** (§4.4): How does the scale size of human scores and metric outputs affect the values of correlation measures?
- **RQ3** (§5.1): Which correlation measures have stronger discriminative power for distinguishing pairs of automatic evaluation metrics?
- **RQ4** (§5.2): Which correlation measures provide more stable rankings for a set of automatic evaluation metrics?

For **RQ1** and **RQ2**, we conduct analyses through statistical simulations. For **RQ3** and **RQ4**, we empirically evaluate various correlation measures using real datasets and real automatic evaluation metrics. Our contributions are summarized as follows:

1) We modeled NLG meta-evaluation and analyzed the relationships between various correlation measures and the impact of scale size on them through large-scale statistical simulations.

2) We designed algorithms for empirically measuring the discriminative power and ranking consistency of correlation measures.

3) We collected the output scores of 32 evaluation metrics across 6 datasets and conducted empirical evaluations of the discriminative power and ranking consistency of various correlation measures.

2 Background

In the field of NLG, a system s takes a source document d as input and generates a target text h. For example, for a news summarization system, the input is news, and the output is a summary. Regarding the target text h generated by the system, we are concerned with its quality. There are two ways to evaluate its quality: human evaluation and automatic evaluation, usually expressed as scores. Human evaluation scores are considered the gold standard, and the consistency between automatic evaluation metrics and human evaluations is used to assess the performance of an automatic evaluation metric m. This process can be formalized as follows:

There are N systems, $\{s_i\}_{i=1}^{N}$ and M source inputs, $\{d_j\}_{j=1}^{M}$. Each source input d_j has corresponding other related content (such as references) v_j. Each system s_i generates a target text h_{ij} for each source input d_j. The human evaluation score for each target text h_{ij} is z_{ij}. The above forms a meta-evaluation dataset $D = \{\{(d_j, v_j)\}_{j=1}^{M}, \{h_{ij}, z_{ij}\}_{i=1,j=1}^{N,M}\}$. In most meta-evaluation datasets, $N \ll M$; generally, the range of N is a few to dozens, while the range of M is tens to thousands.

The input of an automatic evaluation metric m includes a source input d, a target text h, and other related content v, and the output is a score x. For each h_{ij}, the score given by this automatic evaluation metric is denoted as x_{ij}. If there are K automatic evaluation metrics to be evaluated, they are denoted as $\{m_k\}_{k=1}^{K}$, and their output scores are denoted as $\{x_{ij}^k\}_{k=1}^{K}$ (also denoted as matrices $\{X_k\}_{k=1}^{K}$).

The correlation between $\{x_{ij}\}_{i=1,j=1}^{N,M}$ (i.e. X) and $\{z_{ij}\}_{i=1,j=1}^{N,M}$ (i.e. Z) is used to evaluate the quality of the automatic evaluation metrics, and there are multiple ways to measure this correlation, which can be divided into four types based on the grouping method. [1], where c denotes specific correlation coefficient functions, commonly Pearson's r, Spearman's ρ, and Kendall's τ.

- Global level: Calculate the correlation coefficient between two $N \times M$-dimensional vectors, $c_{N\times M}(X, Z) = c(\{(x_{ij}, z_{ij})\}_{i=1,j=1}^{N,M})$.
- Input level: Each time, calculate the correlation coefficient between two N-dimensional vectors, and then average the M correlation coefficients, $c_{\widetilde{N}}(X, Z) = \frac{1}{M}\sum_{j=1}^{M} c(\{(x_{ij}, z_{ij})\}_{i=1}^{N})$. The tilde indicates that this is the average of the correlation coefficients, the same below.
- Item level [2]: Each time, calculate the correlation coefficient between two M-dimensional vectors, and then average the

[1] From a completeness perspective, there is another measure similar to system level, which first averages the scores of each source input across N systems, and then calculates the correlation coefficient between the two M-dimensional vectors. However, this measure may reflect the difficulty of the source inputs and has no significance in evaluation.

[2] WMT22 (Freitag et al., 2022) used this correlation measure as a type of segment-level correlation. We rename it "Item level" to avoid confusion.

N correlation coefficients, $c_{\widetilde{M}}(X, Z) = \frac{1}{N}\sum_{i=1}^{N} c(\{(x_{ij}, z_{ij})\}_{j=1}^{M})$.

- System level: First average the scores of each system across M documents, and then calculate the correlation coefficient between the two N-dimensional vectors, $c_N(X, Z) = c(\{(\frac{1}{M}\sum_{j=1}^{M} x_{ij}, \frac{1}{M}\sum_{j=1}^{M} z_{ij})_{i=1}^{N}\})$.

It can be seen that the measurement of correlation includes two parts: the grouping method and the correlation coefficient function. We use the letters r, ρ, τ to represent the three correlation coefficient functions, and the subscripts $N \times M, \widetilde{N}, \widetilde{M}, N$ to represent the four grouping methods, such as $\tau_{\widetilde{N}}$. Even without considering variants of the correlation coefficient functions, there are $4 \times 3 = 12$ correlation measures.

For two automatic evaluation metrics m_1 and m_2, it is generally considered that m_1 outperforms m_2 if $c(X_1, Z) > c(X_2, Z)$. Here, hypothesis testing can be used to demonstrate statistical significance.

3 Data Preparation

We first obtain the outputs of major automatic evaluation metrics on representative datasets. This not only prepares data for empirical evaluation but also provides references for parameter settings in simulation experiments.

3.1 Datasets

As shown in Table 1, we selected and preprocessed six datasets from five typical NLG tasks: summarization, story generation, dialogue, data-to-text, and translation. Due to the large volume of WMT23 data, we only selected news domain data from ZH2EN. Following conventions, we split the original datasets according to sub-datasets and dimensions, resulting in a total of 30 meta-evaluation datasets, numbered D1-D30, as shown in Table 5.

3.2 Automatic Evaluation Metrics

We selected 14 common non-LLM evaluation metrics, including string-based metrics BLEU (Papineni et al., 2002), ROUGE-(1,2,L) (Lin, 2004), CHRF (Popovic, 2015), and model-based metrics BERTScore-(p,r,f1) (Zhang et al., 2020), MoverScore (Zhao et al., 2019), BARTScore-(s-h, r-h, h-r) (Yuan et al., 2021), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020). For LLM evaluators, we used 18 experimental settings to prompt proprietary LLMs to score target texts based on task descriptions and aspect definitions, resulting in 18 evaluation metrics: three different proprietary LLMs from OpenAI [3] (`gpt-3.5-turbo-1106`, `gpt-4-turbo-1106`, `gpt-4o`); three different scoring scale prompting strategies (1-5, 1-10, 0-100); and two sampling settings (T=0 sampled once, T=1 sampled ten times and averaged). Different scoring scales and sampling settings can significantly change the number of unique values [4] and tie ratio in the metric outputs. More details of the selected evaluation metrics are shown in Appendix A. In total, there are $K = 32$ automatic evaluation metrics.

3.3 Metric Output

For each meta-evaluation dataset, we obtained output scores for all target texts from the 32 metrics. Approximately 0.5% of the LLMs' responses did not score as required. To prevent NAN values from hindering the calculation of various correlation measures, we replaced these with scores randomly sampled from the required scale.

4 Simulation Analysis

Real datasets and evaluation metrics are influenced by multiple factors, making it difficult to control variables. Therefore, this section illustrates through statistical simulations how the correlation between the output of an automatic evaluation metric and human scores is affected by relevant factors under various correlation measures. We consider two factors: the capability of the evaluation metric and the scale size of metric outputs and human scores. We first establish a probabilistic model for NLG evaluation and then obtain results through repeated sampling.

4.1 Modeling NLG Meta-Evaluation

We posit that the capability of an evaluation metric can be decomposed into two parts: the ability to evaluate the overall level of different systems and the ability to evaluate different target texts under a given system. In practice, system-level correlation can estimate the former, while item-level correlation can estimate the latter. Therefore, in our modeling, we treat these two quantities as control parameters. Assuming a given system s_i, the scores of the evaluation metric and human evaluation for texts generated from various

[3] https://openai.com/api/

[4] We refer to this as scale size.

Task	Name	#Subsets	#Aspects	#Systems	#Inputs
Summarization	SummEval (Fabbri et al., 2021)	1	4	16	100
Translation	WMT23-ZH2EN-NEWS (Freitag et al., 2023)	1	1	16	376
Story Generation	HANNA (Chhun et al., 2022)	1	6	6	60
Story Generation	MANS(Guan et al., 2021)	2	1	5	200
Dialogue	USR (Mehri and Eskénazi, 2020)	2	6	5	60
Data-to-text	WebNLG2020 (Castro Ferreira et al., 2020)	1	5	16	178

Table 1: Dataset information.

input documents follow a bivariate normal distribution: $x_{ij}, z_{ij} \sim \mathcal{N}(\mu_i^m, \mu_i^h, \sigma_i^m, \sigma_i^h, \rho_i)$, where ρ_i [5] controls the correlation of the metric within a single system. Based on our observations and those of (Shen et al., 2023), the correlation with human judgment varies across different systems for most evaluation metrics. For simplicity, we assume ρ_i follows a truncated normal distribution, $\rho_i \sim \mathcal{N}(\mu_{\rho_{item}}, \sigma_{\rho_{item}})$. Since item-level correlation is defined as the mean correlation coefficient across different systems, it can be viewed as an estimate of $\mu_{\rho_{item}}$. Furthermore, assuming the parameters μ_i^m and $\mu_i^h$of the above bivariate normal distribution follow another bivariate normal distribution: $\mu_i^m, \mu_i^h \sim \mathcal{N}(\mu^m, \mu^h, \sigma^m, \sigma^h, \rho_{sys})$, where ρ_{sys} controls the correlation between μ_i^m and μ_i^h. system-level correlation can be seen as an estimate of ρ_{sys} because $\frac{1}{M}\sum_{j=1}^{M} x_{ij}$ and $\frac{1}{M}\sum_{j=1}^{M} z_{ij}$ in the definition of system-level correlation are viewed as estimates of μ_i^m and μ_i^h.

Additionally, as mentioned in Section 1, human scores and metric outputs in real-world scenarios cannot always be regarded as fully continuous values. Their degree of discretization can be measured by the number of unique values and tie ratio. For example, SummEval uses a 5-point Likert scale to obtain raw human annotations, with each sample being evaluated by three annotators. After aggregating the scores by averaging, there are up to 13 unique values. For continuous metrics such as BERTScore, they almost never output equal scores across different samples, resulting in a tie ratio close to zero. We have statistically analyzed these two quantities for human scores and metric outputs across different datasets, with results shown in Tables 5 and 6 in the appendix. To simulate different scale sizes, we assume the scale size of human scores and metric outputs to be C^h and C^m respectively. We then follow Onoshima et al. (2019)'s practice to sample $C^h - 1$ and $C^m - 1$ thresholds from uniform distributions $U(-\sigma^m, \sigma^m)$ and $U(-\sigma^h, \sigma^h)$ to discretize them. Algorithm 1 shows the pseudocode for the entire sampling process.

[5]This ρ does not refer to the Spearman correlation coefficient, the same below.

Algorithm 1 Statistical Simulation

Input: $\mu^m, \mu^h, \sigma^m, \sigma^h, \sigma_i^m, \dots, \sigma_N^m, \sigma_1^h, \dots, \sigma_N^h \in \mathbb{R}$, $N, M, C^m, C^h, T_1, T_2 \in \mathbb{N}$, $\rho_{sys}, \mu_{\rho_{item}}, \sigma_{\rho_{item}} \in [-1, 1]$, correlation measure c.
Output: correlation coefficient

$R \leftarrow$ an empty list
for T_1 iterations **do**
  $X^s, Z^s \leftarrow$ empty $N \times M$ matrices
  for $i \in \{1, \dots, N\}$ **do**
    sample $\mu_i^m, \mu_i^h \sim \mathcal{N}(\mu^m, \mu^h, \sigma^m, \sigma^h, \rho_{sys})$
    sample $\rho_i \sim \mathcal{N}(\mu_{\rho_{item}}, \sigma_{\rho_{item}})$
    for $j \in \{1, \dots, M\}$ **do**
      sample $x_{ij}, z_{ij} \sim \mathcal{N}(\mu_i^m, \mu_i^h, \sigma_i^m, \sigma_i^h, \rho_i)$
      $X^s[i, j] \leftarrow x_{ij}$
      $Z^s[i, j] \leftarrow z_{ij}$
    end for
  end for
  if discretization is true **then**
    for T_2 iterations **do**
      sample $\{t_n^m\}_{n=1}^{C^m-1} \sim U(-\sigma^m, \sigma^m)$
      sample $\{t_n^h\}_{n=1}^{C^h-1} \sim U(-\sigma^h, \sigma^h)$
      $X^s \leftarrow \textsc{Discretize}(X^s, \{t_n^m\}_{n=1}^{C^m})$
      $Z^s \leftarrow \textsc{Discretize}(Z^s, \{t_n^h\}_{n=1}^{C^h})$
      $c^s \leftarrow c(X^s, Z^s)$
      Add c^s to R
    end for
  else
    $c^s \leftarrow c(X^s, Z^s)$
    Add c^s to R
  end if
end for
return $\textsc{Avg}(R)$

4.2 Experimental Settings

For the data collected in Section 3, all human scores and metric outputs are normalized to the 0-1 scale for parameter estimation, with results shown in Tables 5 and 6 in the appendix. Balancing the estimated results and without loss of generality, we fix the following parameters for all experiments: $\mu^m = \mu^h = 0$, $\mu_1^m = \dots = \mu_N^m = 0$, $\mu_1^h = \dots = \mu_N^h = 0$, $\sigma^m = \sigma_1^m = \dots = \sigma_N^m = 0.15$, $\sigma^h = \sigma_1^h = \dots = \sigma_N^h = 0.10$, and $\sigma_{\rho_{item}} = 0.15$. For the number of systems and input documents, we consider two settings: $N = 15, M = 200$ and $N = 5, M = 100$.

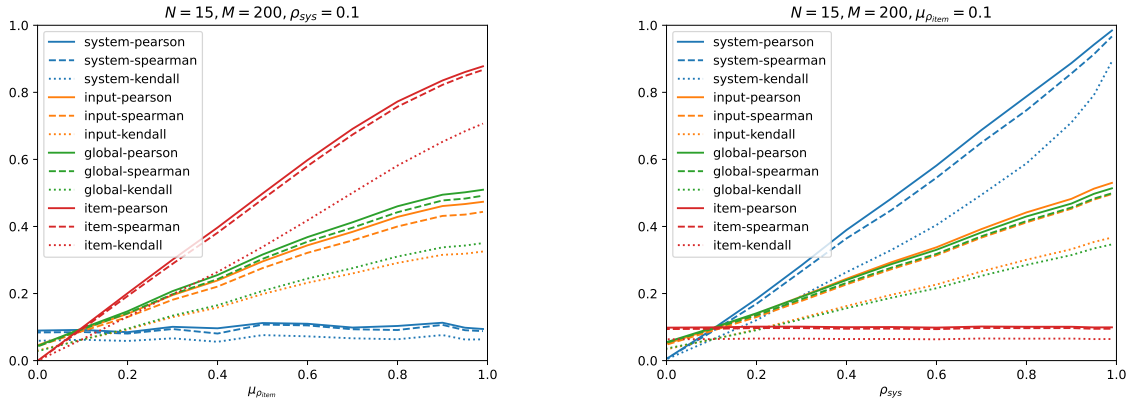

Figure 1: Simulation results of controlling ρ_{sys} and $\mu_{\rho_{item}}$ separately. The result of $N = 5, M = 100$ is shown in Figure 3 in the appendix.

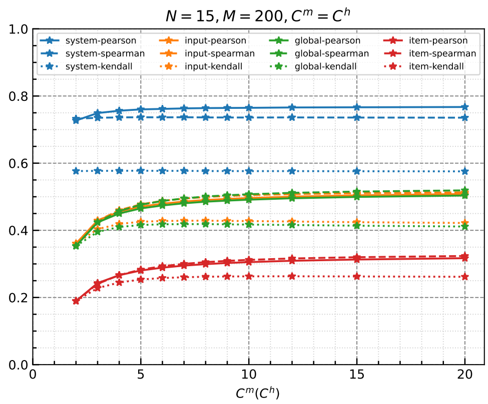

Figure 2: Simulation results of discretization. The result of $N = 5, M = 100$ is shown in Figure 4 in the appendix.

When examining the relationship between different levels of correlation measures, we selected all cases of $\rho_{sys}, \mu_{\rho_{item}} \in \{0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.99\}$, with $T_1 = 1000$ iterations and without discretization.

When analyzing scale size, we fixed $\rho_{sys} = 0.80$ and $\mu_{\rho_{item}} = 0.40$. Regarding C^h and C^m, we considered two different scenarios: for $C^h = C^m$, we selected $C^h = C^m \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20\}$; for $C^m \geq C^h$, we selected all cases satisfying the size relationship from $C^h, C^m \in \{3, 5, 10, 50, 100\}$. Due to the huge amount of computation, we set $T_1 = T_2 = 100$.

4.3 Relationship between Different Levels

In this experiment, we analyze the relationship between different levels of correlation measures. Since we controlled $\mu_{\rho_{item}}$ and ρ_{sys}, it is expected that their estimated values, item-level correlation, and system-level correlation, remain constant or increase. As seen in Figure 1, increasing $\mu_{\rho_{item}}$ or ρ_{sys} results in an increase in input-level correlation and global-level correlation, with both showing similar increments. This trend is also observed in smaller sample scenarios ($N = 5, M = 100$), although the curves are less smooth due to higher sampling variance.

Takeaways Enhancing the evaluation capability of an evaluation metric at the system level or item level will result in higher global-level and input-level correlations.

4.4 Effects of Scale Size

In this experiment, we analyze the effect of different scale sizes on various correlation measures. Figure 2 illustrates that when $C^m = C^h$, i.e., the scale sizes of metric outputs and human scores are equal, increasing the scale size initially increases the values of most correlation measures. These values stabilize after approximately 10, with the Kendall coefficients at the input level, global level, and item level showing a slight decline thereafter. The Spearman and Kendall correlation coefficients at the system level are almost unaffected by scale size.

When $C^m \geq C^h$, the situation is similar: the Spearman and Kendall correlation coefficients at the system level maintain their stability with respect to scale size, as observed when $C^m = C^h$. The Kendall coefficients at the input level, global level, and item level show a central convergence, peaking around $C^m = C^h = 10$. The values of other types of correlation measures increase as C^m

Level	Function	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
System	Pearson	0.074 (5)	0.054 (2)	0.148 (10)	0.150 (10)	0.056 (7)	0.092 (3)	0.035 (1)	0.065 (1)	0.055 (1)	0.125 (7)
System	Spearman	0.137 (9)	0.141 (11)	0.135 (7)	0.174 (11)	0.159 (11)	0.195 (11)	0.174 (9)	0.213 (11)	0.175 (8)	0.209 (11)
System	Kendall	0.178 (12)	0.138 (10)	0.172 (12)	0.228 (12)	0.171 (12)	0.251 (12)	0.214 (12)	0.263 (12)	0.219 (12)	0.246 (12)
Input	Pearson	0.053 (1)	0.064 (3)	0.059 (2)	0.082 (2)	0.077 (9)	0.106 (6)	0.085 (5)	0.144 (7)	0.093 (3)	0.127 (8)
Input	Spearman	0.074 (4)	0.086 (6)	0.096 (5)	0.103 (6)	0.085 (10)	0.117 (7)	0.097 (6)	0.121 (5)	0.117 (7)	0.096 (2)
Input	Kendall	0.074 (6)	0.070 (4)	0.099 (6)	0.107 (8)	0.061 (8)	0.105 (5)	0.101 (7)	0.122 (6)	0.116 (6)	0.107 (4)
Global	Pearson	0.064 (2)	0.054 (1)	0.058 (1)	0.079 (1)	0.035 (3)	0.089 (2)	0.070 (3)	0.121 (4)	0.085 (2)	0.084 (1)
Global	Spearman	0.071 (3)	0.114 (8)	0.090 (3)	0.100 (5)	0.023 (1)	0.095 (4)	0.067 (2)	0.095 (2)	0.100 (5)	0.100 (3)
Global	Kendall	0.085 (7)	0.084 (5)	0.096 (4)	0.097 (4)	0.053 (6)	0.084 (1)	0.079 (4)	0.115 (3)	0.093 (4)	0.117 (5)
Item	Pearson	0.135 (8)	0.102 (7)	0.146 (9)	0.094 (3)	0.038 (4)	0.138 (8)	0.162 (8)	0.175 (8)	0.175 (9)	0.122 (6)
Item	Spearman	0.147 (10)	0.134 (9)	0.149 (11)	0.106 (7)	0.027 (2)	0.141 (9)	0.180 (10)	0.177 (9)	0.179 (11)	0.132 (9)
Item	Kendall	0.151 (11)	0.147 (12)	0.145 (8)	0.108 (9)	0.041 (5)	0.147 (10)	0.206 (11)	0.190 (10)	0.179 (10)	0.148 (10)

Level	Function	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
System	Pearson	0.062 (1)	0.214 (10)	0.170 (7)	0.171 (5)	0.162 (10)	0.147 (4)	0.244 (10)	0.451 (10)	0.209 (1)	0.101 (4)
System	Spearman	0.188 (8)	0.433 (11)	0.424 (11)	0.374 (11)	0.355 (11)	0.359 (11)	0.375 (11)	0.544 (11)	0.349 (3)	0.579 (11)
System	Kendall	0.225 (12)	0.444 (12)	0.505 (12)	0.386 (12)	0.358 (12)	0.374 (12)	0.396 (12)	0.623 (12)	0.361 (4)	0.651 (12)
Input	Pearson	0.135 (5)	0.102 (4)	0.135 (4)	0.167 (4)	0.119 (3)	0.170 (5)	0.116 (3)	0.351 (7)	0.336 (2)	0.092 (2)
Input	Spearman	0.123 (4)	0.114 (6)	0.141 (5)	0.182 (6)	0.145 (7)	0.186 (6)	0.158 (7)	0.392 (9)	0.414 (5)	0.138 (6)
Input	Kendall	0.151 (7)	0.106 (5)	0.145 (6)	0.189 (7)	0.157 (8)	0.194 (9)	0.164 (8)	0.373 (8)	0.414 (6)	0.143 (7)
Global	Pearson	0.136 (6)	0.072 (1)	0.107 (1)	0.093 (1)	0.127 (5)	0.108 (1)	0.088 (1)	0.165 (1)	0.540 (9)	0.080 (1)
Global	Spearman	0.105 (2)	0.072 (2)	0.116 (3)	0.134 (3)	0.094 (1)	0.121 (3)	0.092 (2)	0.219 (3)	0.487 (8)	0.101 (3)
Global	Kendall	0.107 (3)	0.095 (3)	0.110 (2)	0.129 (2)	0.110 (2)	0.119 (2)	0.128 (4)	0.193 (2)	0.483 (7)	0.104 (5)
Item	Pearson	0.201 (9)	0.148 (7)	0.192 (9)	0.215 (8)	0.127 (4)	0.192 (8)	0.148 (5)	0.232 (4)	0.569 (10)	0.192 (8)
Item	Spearman	0.214 (10)	0.152 (8)	0.186 (8)	0.224 (9)	0.140 (6)	0.192 (7)	0.152 (6)	0.265 (6)	0.603 (12)	0.227 (9)
Item	Kendall	0.219 (11)	0.182 (9)	0.196 (10)	0.234 (10)	0.157 (9)	0.210 (10)	0.180 (9)	0.262 (5)	0.601 (11)	0.232 (10)

Level	Function	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30
System	Pearson	0.157 (7)	0.140 (6)	0.134 (6)	0.149 (7)	0.159 (1)	0.092 (3)	0.076 (2)	0.155 (10)	0.109 (3)	0.127 (10)
System	Spearman	0.563 (11)	0.616 (11)	0.642 (11)	0.573 (11)	0.602 (11)	0.262 (11)	0.155 (10)	0.314 (11)	0.472 (11)	0.282 (11)
System	Kendall	0.595 (12)	0.669 (12)	0.711 (12)	0.622 (12)	0.698 (12)	0.290 (12)	0.173 (12)	0.408 (12)	0.532 (12)	0.337 (12)
Input	Pearson	0.117 (4)	0.110 (3)	0.106 (3)	0.107 (1)	0.263 (6)	0.110 (6)	0.115 (3)	0.080 (3)	0.123 (4)	0.092 (5)
Input	Spearman	0.158 (8)	0.136 (4)	0.123 (5)	0.147 (6)	0.271 (7)	0.138 (10)	0.145 (9)	0.101 (8)	0.184 (10)	0.125 (9)
Input	Kendall	0.167 (10)	0.140 (5)	0.135 (7)	0.145 (5)	0.259 (5)	0.128 (8)	0.129 (6)	0.084 (5)	0.174 (9)	0.106 (8)
Global	Pearson	0.117 (3)	0.096 (1)	0.085 (1)	0.114 (2)	0.225 (3)	0.064 (1)	0.064 (1)	0.054 (1)	0.074 (1)	0.061 (2)
Global	Spearman	0.106 (2)	0.104 (2)	0.104 (2)	0.117 (3)	0.233 (4)	0.088 (2)	0.120 (5)	0.068 (2)	0.135 (5)	0.058 (1)
Global	Kendall	0.096 (1)	0.153 (7)	0.109 (4)	0.118 (4)	0.221 (2)	0.107 (4)	0.140 (8)	0.093 (7)	0.155 (6)	0.077 (3)
Item	Pearson	0.129 (5)	0.168 (8)	0.178 (9)	0.193 (10)	0.332 (8)	0.127 (7)	0.118 (4)	0.083 (4)	0.108 (2)	0.095 (6)
Item	Spearman	0.164 (9)	0.197 (9)	0.178 (8)	0.180 (8)	0.364 (9)	0.108 (5)	0.134 (7)	0.091 (6)	0.171 (7)	0.087 (4)
Item	Kendall	0.157 (6)	0.247 (10)	0.182 (10)	0.191 (9)	0.367 (10)	0.135 (9)	0.163 (11)	0.110 (9)	0.174 (8)	0.104 (7)

Table 2: DP values of different correlation measures on all meta-evaluation datasets using permutation test, the lower the better. Each column "DN" shows the result on a meta-evaluation dataset, and the mapping to the original dataset is shown in Table 5 in the appendix. The results using William's test are shown in Table 7 in the appendix.

or C^h increases, possibly stabilizing after reaching a certain point. Specific values are detailed in Figures 5 and 6 in the appendix.

Takeaways 1) The values of system-level Spearman's ρ and Kendall's τ are almost unaffected by scale size. 2) For input-level, global-level, and item-level Kendall's τ, the effect of scale size is complex; as the scale size of human scores or metric outputs increases, their values first rise and then fall. 3) As the scale size of human scores or metric outputs increases, the values of other correlation measures increase.

5 Empirical Evaluation

Through the above simulation analysis, we better understand how the correlation between an automatic evaluation metric and human evaluation is affected by relevant factors. However, in practice, correlation measures are mainly used to compare the performance of different evaluation metrics, including two primary uses: comparing the performance of two automatic evaluation metrics and ranking the performance of a set of automatic evaluation metrics. For the former, we assess the discriminative power of the correlation measure, i.e., whether it can distinguish as many pairs of metrics as possible. For the latter, we evaluate the consistency of the correlation measure in ranking evaluation metrics, i.e., whether the ranking is stable.

5.1 Discriminative Power

In the fields of information retrieval (Sakai, 2013) and recommendation systems (Anelli et al., 2019; Ashkan and Metzler, 2019; Valcarce et al., 2020), Discriminative Power is widely used to compare evaluation measures. Inspired by this, we adapted this method to evaluate correlation measures in NLG meta-evaluation.

Specifically, for a given correlation measure, a meta-evaluation dataset (including human scores Z), and the scores of K automatic evaluation metrics on it $\{X_k\}_{k=1}^K$, we obtain the two-sided p-value for each pair of automatic evaluation metrics through hypothesis testing. The smaller the p-value, the more confidently we can claim that the two correlations differ. A highly discriminative correlation measure will yield many very small p-

Level	Function	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
System	Pearson	0.789 (9)	0.726 (12)	0.884 (5)	0.707 (10)	0.911 (2)	0.918 (4)	0.918 (3)	0.954 (1)	0.950 (2)	0.912 (5)
System	Spearman	0.758 (10)	0.801 (11)	0.780 (10)	0.686 (11)	0.679 (12)	0.627 (11)	0.707 (12)	0.584 (12)	0.707 (11)	0.643 (11)
System	Kendall	0.729 (12)	0.810 (10)	0.784 (9)	0.666 (12)	0.702 (11)	0.588 (12)	0.708 (11)	0.585 (11)	0.685 (12)	0.631 (12)
Input	Pearson	0.890 (3)	0.949 (1)	0.944 (2)	0.857 (6)	0.954 (1)	0.972 (1)	0.942 (2)	0.950 (2)	0.954 (1)	0.953 (1)
Input	Spearman	0.853 (6)	0.893 (5)	0.787 (8)	0.856 (7)	0.790 (10)	0.879 (7)	0.805 (8)	0.892 (8)	0.808 (8)	0.909 (6)
Input	Kendall	0.901 (1)	0.912 (4)	0.909 (4)	0.860 (4)	0.905 (3)	0.870 (9)	0.866 (6)	0.936 (5)	0.851 (7)	0.943 (3)
Global	Pearson	0.852 (7)	0.940 (3)	0.952 (1)	0.868 (2)	0.877 (6)	0.970 (2)	0.958 (1)	0.942 (3)	0.936 (3)	0.943 (2)
Global	Spearman	0.857 (4)	0.835 (8)	0.754 (11)	0.853 (8)	0.835 (9)	0.874 (8)	0.795 (9)	0.885 (9)	0.803 (10)	0.896 (7)
Global	Kendall	0.891 (2)	0.872 (7)	0.803 (7)	0.864 (3)	0.841 (8)	0.915 (5)	0.872 (5)	0.938 (4)	0.863 (6)	0.937 (4)
Item	Pearson	0.834 (8)	0.941 (2)	0.937 (3)	0.896 (1)	0.898 (4)	0.886 (6)	0.861 (7)	0.912 (6)	0.905 (4)	0.842 (10)
Item	Spearman	0.755 (11)	0.814 (9)	0.706 (12)	0.792 (9)	0.851 (7)	0.856 (10)	0.776 (10)	0.881 (10)	0.805 (9)	0.891 (9)
Item	Kendall	0.854 (5)	0.889 (6)	0.861 (6)	0.858 (5)	0.880 (5)	0.920 (3)	0.895 (4)	0.894 (7)	0.899 (5)	0.893 (8)
Level	Function	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
System	Pearson	0.936 (3)	0.898 (2)	0.891 (5)	0.897 (5)	0.644 (12)	0.888 (3)	0.767 (11)	0.859 (7)	0.884 (1)	0.845 (4)
System	Spearman	0.615 (11)	0.845 (7)	0.595 (12)	0.939 (1)	0.698 (11)	0.676 (12)	0.805 (8)	0.517 (11)	0.692 (5)	0.322 (12)
System	Kendall	0.605 (12)	0.846 (6)	0.647 (11)	0.938 (2)	0.700 (10)	0.719 (11)	0.805 (7)	0.488 (12)	0.702 (4)	0.354 (11)
Input	Pearson	0.941 (1)	0.892 (3)	0.919 (2)	0.899 (4)	0.917 (1)	0.915 (1)	0.829 (5)	0.926 (2)	0.585 (7)	0.796 (5)
Input	Spearman	0.915 (7)	0.771 (11)	0.871 (9)	0.656 (12)	0.769 (9)	0.784 (9)	0.788 (10)	0.640 (10)	0.533 (9)	0.712 (10)
Input	Kendall	0.918 (6)	0.850 (5)	0.904 (3)	0.777 (8)	0.850 (3)	0.822 (8)	0.789 (9)	0.867 (6)	0.529 (11)	0.776 (6)
Global	Pearson	0.926 (4)	0.802 (9)	0.875 (6)	0.907 (3)	0.839 (4)	0.894 (2)	0.916 (1)	0.948 (1)	0.787 (3)	0.860 (1)
Global	Spearman	0.923 (5)	0.733 (12)	0.875 (8)	0.697 (10)	0.790 (6)	0.846 (7)	0.866 (4)	0.738 (9)	0.533 (10)	0.746 (8)
Global	Kendall	0.941 (2)	0.820 (8)	0.920 (1)	0.752 (9)	0.829 (5)	0.868 (5)	0.895 (2)	0.900 (4)	0.622 (6)	0.762 (7)
Item	Pearson	0.912 (8)	0.790 (10)	0.875 (7)	0.890 (6)	0.770 (8)	0.846 (6)	0.875 (3)	0.924 (3)	0.789 (2)	0.859 (2)
Item	Spearman	0.891 (9)	0.863 (4)	0.829 (10)	0.663 (11)	0.790 (7)	0.751 (10)	0.758 (12)	0.771 (8)	0.362 (12)	0.729 (9)
Item	Kendall	0.887 (10)	0.913 (1)	0.903 (4)	0.846 (7)	0.869 (2)	0.871 (4)	0.828 (6)	0.887 (5)	0.541 (8)	0.857 (3)
Level	Function	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30
System	Pearson	0.853 (4)	0.838 (5)	0.720 (10)	0.790 (10)	0.766 (1)	0.927 (1)	0.831 (8)	0.877 (6)	0.857 (7)	0.912 (4)
System	Spearman	0.339 (11)	0.344 (12)	0.274 (12)	0.282 (12)	0.338 (12)	0.451 (12)	0.748 (11)	0.521 (12)	0.233 (12)	0.527 (12)
System	Kendall	0.331 (12)	0.365 (11)	0.298 (11)	0.298 (11)	0.344 (11)	0.520 (11)	0.774 (10)	0.547 (11)	0.255 (11)	0.556 (11)
Input	Pearson	0.903 (1)	0.885 (3)	0.869 (3)	0.917 (1)	0.599 (8)	0.910 (3)	0.920 (1)	0.882 (5)	0.909 (1)	0.933 (1)
Input	Spearman	0.775 (9)	0.648 (9)	0.760 (7)	0.792 (9)	0.546 (9)	0.818 (7)	0.851 (7)	0.783 (9)	0.838 (8)	0.736 (9)
Input	Kendall	0.881 (2)	0.790 (6)	0.801 (5)	0.837 (5)	0.646 (6)	0.808 (8)	0.856 (6)	0.889 (4)	0.877 (5)	0.859 (6)
Global	Pearson	0.881 (3)	0.905 (2)	0.889 (2)	0.902 (2)	0.751 (2)	0.907 (4)	0.919 (2)	0.890 (3)	0.897 (3)	0.931 (2)
Global	Spearman	0.782 (8)	0.636 (10)	0.756 (8)	0.832 (6)	0.621 (7)	0.678 (10)	0.703 (12)	0.681 (10)	0.707 (10)	0.691 (10)
Global	Kendall	0.819 (7)	0.695 (7)	0.761 (6)	0.865 (4)	0.727 (3)	0.880 (6)	0.886 (5)	0.833 (7)	0.877 (6)	0.769 (8)
Item	Pearson	0.835 (6)	0.907 (1)	0.890 (1)	0.867 (3)	0.705 (4)	0.922 (2)	0.913 (3)	0.902 (1)	0.902 (2)	0.923 (3)
Item	Spearman	0.763 (10)	0.688 (8)	0.721 (9)	0.810 (8)	0.506 (10)	0.757 (9)	0.806 (9)	0.795 (8)	0.814 (9)	0.815 (7)
Item	Kendall	0.838 (5)	0.869 (4)	0.833 (4)	0.830 (7)	0.646 (5)	0.905 (5)	0.901 (4)	0.898 (2)	0.882 (4)	0.880 (5)

Table 3: RC values of different correlation measures on all meta-evaluation datasets. Each column "DN" shows the result on a meta-evaluation dataset, and the mapping to the original dataset is shown in Table 5 in the appendix.

values. After obtaining the p-values for each pair of evaluation metrics, we sort them in descending order. With the number of evaluation metric pairs on the x-axis and the p-values on the y-axis, we can plot the p-value curves of different correlation measures on a meta-evaluation dataset. The closer the curve is to the coordinate axis, the stronger the Discriminative Power of the corresponding correlation measure. For convenience of comparison, similar to Valcarce et al. (2020), we define the DP value as the average of all p-values, ranging from 0 to 1, with smaller values indicating stronger discriminative power. This value numerically equals the area enclosed by the p-value curve and the coordinate axis, normalized by the number of evaluation metric pairs. Algorithm 2 shows the pseudocode for calculating the DP value.

Regarding the hypothesis testing methods used here, we refer to previous work and employ William's test (Williams, 1959) and permutation test [6] (Noreen, 1989). The former has been proposed for comparing machine translation evaluation metrics (Graham and Baldwin, 2014), and the latter is a non-parametric test method that Deutsch et al. (2021) has shown to have a higher power in summarization meta-evaluation.

Algorithm 2 Discriminative Power

Input: $X_1, \ldots, X_K, Z \in \mathbb{R}^{N \times M}$, correlation measure c.
Output: DP value

```
v ← 0
n ← K × (K − 1)/2
for i ∈ {1, ..., K − 1} do
    for j ∈ {i, ..., K} do
        p_ij ← HypoTest(X_i, X_j, Z, c)
        v ← v + p_ij
    end for
end for
return v/n
```

Table 2 shows the DP values of correlation measures across all meta-evaluation datasets and The p-values curves are shown in Figure 7-36 in the appendix.

Takeaways Despite variations in results across different datasets, the discriminative power of different correlation measures can be summarized as follows:

- Level: Global > Input > Item > System

[6] We use the Perm-Both algorithm proposed by Deutsch et al. (2021), with a sample size of 1000.

- Function: Pearson's r > Spearman's ρ > Kendall's τ

5.2 Ranking Consistency

Inspired by Sakai (2021)'s evaluation of ordinal classification tasks, for a given correlation measure, we randomly split the human scores and evaluation metric outputs in half, derive the rankings of the evaluation metrics on the two halves, and calculate the similarity of the two rankings using τ_b as a measure of ranking consistency. We define the RC value as the mean obtained from repeating this process $T = 1000$ times. Algorithm 3 presents the pseudocode for the calculation. Table 3 shows the RC values of correlation measures across all meta-evaluation datasets.

Takeaways Despite variations in results across different datasets, the ranking consistency of different correlation measures can be summarized as follows:

- Level: Input ≈ Global > Item > System
- Function: Pearson's r > Kendall's τ > Spearman's ρ

Algorithm 3 Ranking Consistency

Input: $X_1, \ldots, X_K, Z \in \mathbb{R}^{N \times M}, T \in \mathbb{N}$, correlation measure c.
Output: RC value

$v \leftarrow 0$
for T iterations **do**
  $M_1 \leftarrow \lfloor M/2 \rfloor$
  $M_2 \leftarrow M - \lfloor M/2 \rfloor$
  $D_1 \leftarrow$ sample $\{1, \ldots, M\}$ w/o repl. M_1 times
  $D_2 \leftarrow \{1, \ldots, M\} \setminus D_1$
  $R_1, R_2 \leftarrow$ empty K-dimensional arrays
  for $k \in \{1, \ldots, K\}$ **do**
    $X_1^s, Z_1^s \leftarrow$ empty $N \times M_1$ matrices
    $X_2^s, Z_2^s \leftarrow$ empty $N \times M_2$ matrices
    for $i \in \{1, \ldots, N\}$ **do**
      for $j \in \{1, \ldots, M_1\}$ **do**
        $X_1^s[i, j] \leftarrow X_k[i, D_1[j]]$
        $Z_1^s[i, j] \leftarrow Z[i, D_1[j]]$
      end for
      for $j \in \{1, \ldots, M_2\}$ **do**
        $X_2^s[i, j] \leftarrow X_k[i, D_2[j]]$
        $Z_2^s[i, j] \leftarrow Z[i, D_2[j]]$
      end for
    end for
    $R_1[k] \leftarrow c(X_1^s, Z_1^s)$
    $R_2[k] \leftarrow c(X_2^s, Z_2^s)$
  end for
  $\tau^s \leftarrow \tau_b(R_1, R_2)$
  $v \leftarrow v + \tau^s$
end for
return v/T

6 Related Work

In the field of NLP, there is limited research analyzing correlation measures in NLG meta-evaluation. Mathur et al. (2020) found that the introduction of an outlier system can distort the system-level Pearson correlation between machine translation evaluation metrics and human evaluation and quantify the impact. Recently, Deutsch et al. (2023) pointed out that the segment-level Kendall correlation coefficient, widely used in machine translation evaluation, does not handle ties in human scores and metric outputs as expected and thus needs to be calibrated. In the study of automatic evaluation metrics, some works have commented on correlation measures based on experimental results when presenting the performance of different evaluation metrics. Owczarzak et al. (2012) found that, in the domain of summarization evaluation, system-level correlation is more robust to inconsistent human annotations. Freitag et al. (2022) discovered that system-level correlation is hard to distinguish between different machine translation evaluation metrics. Liu et al. (2023); Xu et al. (2023) explained some experimental results as the inappropriate handling of ties by the Kendall correlation coefficient when comparing the performance of different metrics. In contrast, we focus on the properties and capabilities of typical correlation measures from a generic NLG evaluation perspective, not limited to specific tasks and evaluation metrics.

7 Conclusion

We analyzed and evaluated the characteristics and capabilities of 12 typical correlation measures in NLG meta-evaluation through statistical simulations and empirical experiments. Based on our experiments, we emphasize the following points:

1) In most current NLG evaluation datasets, system-level correlation measures have the lowest discriminative power and ranking consistency for evaluation metrics. However, system-level Spearman's ρ and Kendall's τ have the advantage of being unaffected by scale size.

2) The most widely used input-level and global-level correlation measures currently exhibit good discriminative power and ranking consistency, but both are influenced by scale size. We call for further research to address this issue.

Limitations

Although our empirical experiments covered many datasets, it is impossible to encompass all tasks and evaluation aspects. Therefore, the conclusions we obtained regarding the discriminative power and ranking consistency of various correlation measures may not be applicable to other tasks and scenarios. Additionally, our work requires substantial funding and computational resources: using GPT-3.5 and GPT-4 to annotate large amounts of data incurred significant costs; conducting large-scale statistical simulations and empirical evaluations required high-performance computing resources, which may hinder others from replicating our work.

References

Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Claudio Pomo, and Azzurra Ragone. 2019. On the discriminative power of hyper-parameters in cross-validation and how to choose them. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, pages 447–451. ACM.

Azin Ashkan and Donald Metzler. 2019. Revisiting online personal search metrics with the user in mind. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 625–634. ACM.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9347–9359. Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina, editors. 2020. *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*. Association for Computational Linguistics, Dublin, Ireland (Virtual).

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5794–5836. International Committee on Computational Linguistics.

David Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15607–15631. Association for Computational Linguistics.

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

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

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George F. 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, WMT 2023, Singapore, December 6-7, 2023*, pages 578–628. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George F. 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 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 46–68. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 172–176. ACL.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual*

Event, August 1-6, 2021, pages 6394–6407. Association for Computational Linguistics.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *CoRR*, abs/2310.00752.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 193–203. European Association for Machine Translation.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *CoRR*, abs/2310.05470.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4984–4997. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskénazi. 2020. USR: an unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 681–707. Association for Computational Linguistics.

Eric W Noreen. 1989. Computer intensive methods for hypothesis testing: An introduction. *Wiley, New York*, 19:21.

Takahiro Onoshima, Kenpei Shiina, Takashi Ueda, and Saori Kubo. 2019. Decline of pearson'sr with categorization of variables: a large-scale simulation. *Behaviormetrika*, 46:389–399.

Karolina Owczarzak, Peter A. Rankel, Hoa Trang Dang, and John M. Conroy. 2012. Assessing the effect of inconsistent assessors on summarization evaluation. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 359–362. The Association for Computer Linguistics.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer 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 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Tetsuya Sakai. 2013. Metrics, statistics, tests. In *Bridging Between Information Retrieval and Databases - PROMISE Winter School 2013, Bressanone, Italy, February 4-8, 2013. Revised Tutorial Lectures*, volume 8173 of *Lecture Notes in Computer Science*, pages 116–163. Springer.

Tetsuya Sakai. 2021. Evaluating evaluation measures for ordinal classification and ordinal quantification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2759–2769. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4215–4233. Association for Computational Linguistics.

Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2020. Assessing ranking metrics in top-n recommendation. *Inf. Retr. J.*, 23(4):411–448.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023. Automated evaluation of personalized text generation using large language models. *CoRR*, abs/2310.11593.

Evan James Williams. 1959. *Regression Analysis*, volume 14. wiley.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5967–5994. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 563–578. Association for Computational Linguistics.

A Details of Selected Evaluation Metrics

A.1 Non-LLM evaluation metrics

For CHRF and BLEU, we use the implementation of TorchMetrics [7]. For ROUGE, BERTSCORE, and BLEURT, we use the `evaluation` package of Huggingface with the default parameters. For MoverScore [8], BARTScore [9], and COMET [10], we use the code from the original GitHub repositories and the default models. We check the licenses of all open source programs to ensure that our use is compliant.

A.2 Evaluation Prompts for LLMs

We used the same prompts to instruct GPT-3.5, GPT-4, and GPT-4o for NLG evaluation. To save space, we present a template of our prompt in Table 4. We filled the aspect part of the prompt with definitions from original datasets. When the original dataset lacked these definitions, we composed them based on our understanding.

[7] https://lightning.ai/docs/torchmetrics/stable/

[8] https://github.com/AIPHES/emnlp19-moverscore

[9] https://github.com/neulab/BARTScore

[10] https://github.com/Unbabel/COMET

B Other Figures and Tables

Prompts and Instructions
###Instruction### Please act as an impartial and helpful evaluator for natural language generation (NLG), and the audience is an expert in the field. Your task is to evaluate the quality of {task} strictly based on the given evaluation criterion. Begin the evaluation by providing your analysis concisely and accurately, and then on the next line, start with "Rating:" followed by your rating on a Likert scale from {scale} (higher means better). You MUST keep to the strict boundaries of the evaluation criterion and focus solely on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion and example content, carefully. ###Evaluation Criterion### {aspect} ###Example### {source_des}: {source} {target_des}: {target} ###Your Evaluation###

Table 4: Prompts and instructions used for LLMs to evaluate and annotate NLG tasks.

Dataset	Subset	Aspect	No.	$\widehat{\mu^h}$	$\widehat{\sigma^h}$	#Unique Values (Human)	Tie Ratio (Human)
SummEval	CNN/DM	Coherence	D1	0.60	0.15	13	0.10
SummEval	CNN/DM	Consistency	D2	0.92	0.14	12	0.67
SummEval	CNN/DM	Fluency	D3	0.92	0.09	13	0.53
SummEval	CNN/DM	Relevance	D4	0.69	0.09	13	0.13
WMT23	GeneralMT2023_NEWS	Overall Quality	D5	0.84	0.05	239	0.11
HANNA	WP	Coherence	D6	0.54	0.13	13	0.13
HANNA	WP	Complexity	D7	0.36	0.14	13	0.12
HANNA	WP	Empathy	D8	0.32	0.09	12	0.13
HANNA	WP	Engagement	D9	0.42	0.13	13	0.12
HANNA	WP	Relevance	D10	0.41	0.14	13	0.10
HANNA	WP	Surprise	D11	0.28	0.10	12	0.15
MANS	ROC	Overall	D12	0.38	0.15	21	0.06
MANS	WP	Overall	D13	0.45	0.08	21	0.08
USR	Persona-Chat	Engaging	D14	0.38	0.07	7	0.24
USR	Persona-Chat	Maintains Context	D15	0.37	0.09	7	0.31
USR	Persona-Chat	Natural	D16	0.44	0.04	7	0.48
USR	Persona-Chat	Overall	D17	0.69	0.19	12	0.11
USR	Persona-Chat	Understandable	D18	-0.01	0.01	4	0.84
USR	Persona-Chat	Uses Knowledge	D19	-0.14	0.09	4	0.38
USR	Topical-Chat	Engaging	D20	0.28	0.13	7	0.15
USR	Topical-Chat	Maintains Context	D21	0.31	0.12	7	0.17
USR	Topical-Chat	Natural	D22	0.32	0.11	7	0.17
USR	Topical-Chat	Overall	D23	0.54	0.27	13	0.08
USR	Topical-Chat	Understandable	D24	-0.08	0.06	4	0.33
USR	Topical-Chat	Uses Knowledge	D25	-0.11	0.06	4	0.33
WebNLG2020	WebNLG2020	Correctness	D26	0.88	0.07	268	0.03
WebNLG2020	WebNLG2020	Datacoverage	D27	0.9	0.06	246	0.04
WebNLG2020	WebNLG2020	Fluency	D28	0.83	0.06	282	0.01
WebNLG2020	WebNLG2020	Relevance	D29	0.91	0.05	226	0.05
WebNLG2020	WebNLG2020	Textstructure	D30	0.87	0.05	247	0.02

Table 5: Divided meta-evaluation dataset information and estimated parameters.

Metric Name	$\widehat{\mu^m}$	$\widehat{\sigma^m}$	$\widehat{\rho_{sys}}$	$\widehat{\mu_{\rho_{item}}}$	$\widehat{\sigma_{\rho_{item}}}$	$r_{\tilde{N}}$	$r_{N\times M}$	#Unique Values	Tie Ratio
GPT3.5_T=0_0_100	0.44	0.08	0.88	0.22	0.13	0.41	0.38	18	0.35
GPT3.5_T=0_1_10	0.33	0.08	0.83	0.20	0.13	0.41	0.35	8	0.46
GPT3.5_T=0_1_5	0.32	0.10	0.86	0.21	0.12	0.40	0.35	5	0.49
GPT3.5_T=1_0_100	0.45	0.08	0.89	0.27	0.14	0.48	0.44	175	0.03
GPT3.5_T=1_1_10	0.33	0.08	0.88	0.26	0.13	0.46	0.41	87	0.10
GPT3.5_T=1_1_5	0.32	0.09	0.88	0.25	0.13	0.46	0.41	46	0.14
GPT4_T=0_0_100	0.51	0.15	0.92	0.35	0.14	0.56	0.55	19	0.22
GPT4_T=0_1_10	0.45	0.15	0.92	0.34	0.13	0.55	0.53	9	0.28
GPT4_T=0_1_5	0.44	0.18	0.92	0.31	0.14	0.55	0.51	5	0.45
GPT4_T=1_0_100	0.51	0.14	0.93	0.39	0.14	0.59	0.58	279	0.02
GPT4_T=1_1_10	0.45	0.15	0.93	0.38	0.13	0.59	0.56	87	0.06
GPT4_T=1_1_5	0.43	0.17	0.92	0.36	0.14	0.58	0.55	44	0.18
GPT4o_T=0_0_100	0.48	0.14	0.92	0.37	0.13	0.57	0.55	20	0.21
GPT4o_T=0_1_10	0.42	0.15	0.91	0.35	0.12	0.55	0.53	9	0.27
GPT4o_T=0_1_5	0.40	0.17	0.91	0.32	0.13	0.54	0.50	5	0.43
GPT4o_T=1_0_100	0.48	0.14	0.92	0.39	0.13	0.59	0.58	272	0.02
GPT4o_T=1_1_10	0.42	0.14	0.92	0.38	0.13	0.59	0.56	81	0.05
GPT4o_T=1_1_5	0.40	0.16	0.91	0.35	0.13	0.57	0.54	46	0.16
BERTScore-f	0.87	0.04	0.60	-0.02	0.10	0.26	0.22	1136	0.01
BERTScore-p	0.87	0.04	0.53	-0.02	0.10	0.23	0.20	1136	0.01
BERTScore-r	0.88	0.04	0.67	-0.01	0.11	0.29	0.24	1136	0.01
BLEU	0.15	0.24	0.52	-0.01	0.09	0.24	0.21	412	0.53
CHRF	0.37	0.20	0.58	0.00	0.10	0.27	0.24	1130	0.02
COMET	0.60	0.12	0.73	-0.01	0.11	0.32	0.27	1190	0.00
MoverScore	0.61	0.11	0.54	-0.02	0.11	0.24	0.22	1180	0.00
ROUGE-1	0.41	0.20	0.53	-0.02	0.11	0.25	0.22	619	0.02
ROUGE-2	0.23	0.23	0.54	-0.01	0.10	0.25	0.22	569	0.15
ROUGE-L	0.33	0.21	0.53	-0.03	0.10	0.23	0.20	623	0.02

Table 6: Metrics information and estimated parameters averaged across meta-evaluation datasets. We do not use the output of BLEURT and BARTScore-(s-h, r-h, h-r) to estimate the parameters because their scores do not have clear ranges.

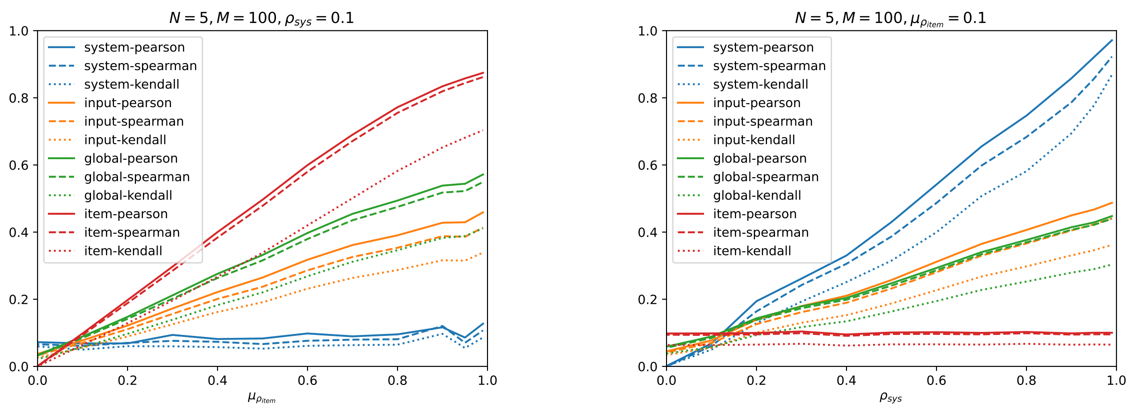

Figure 3: Simulation results of controlling ρ_{sys} and $\mu_{\rho_{item}}$ separately.

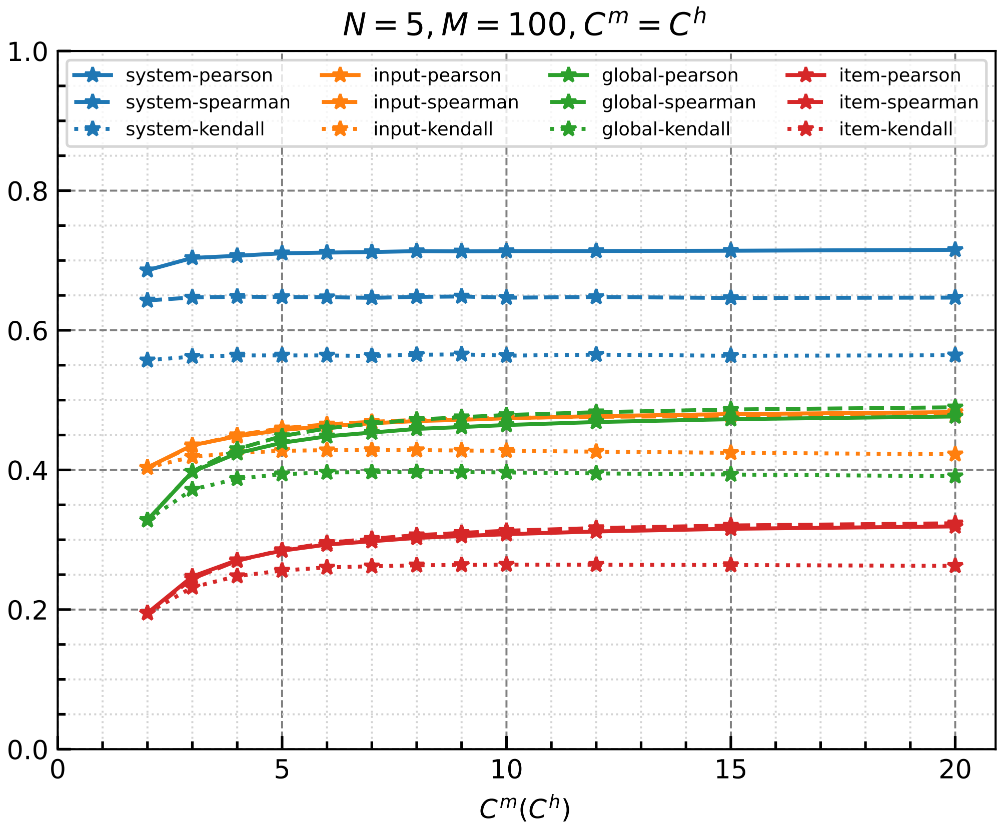

Figure 4: Simulation results of discretization when $C^m = C^h$.

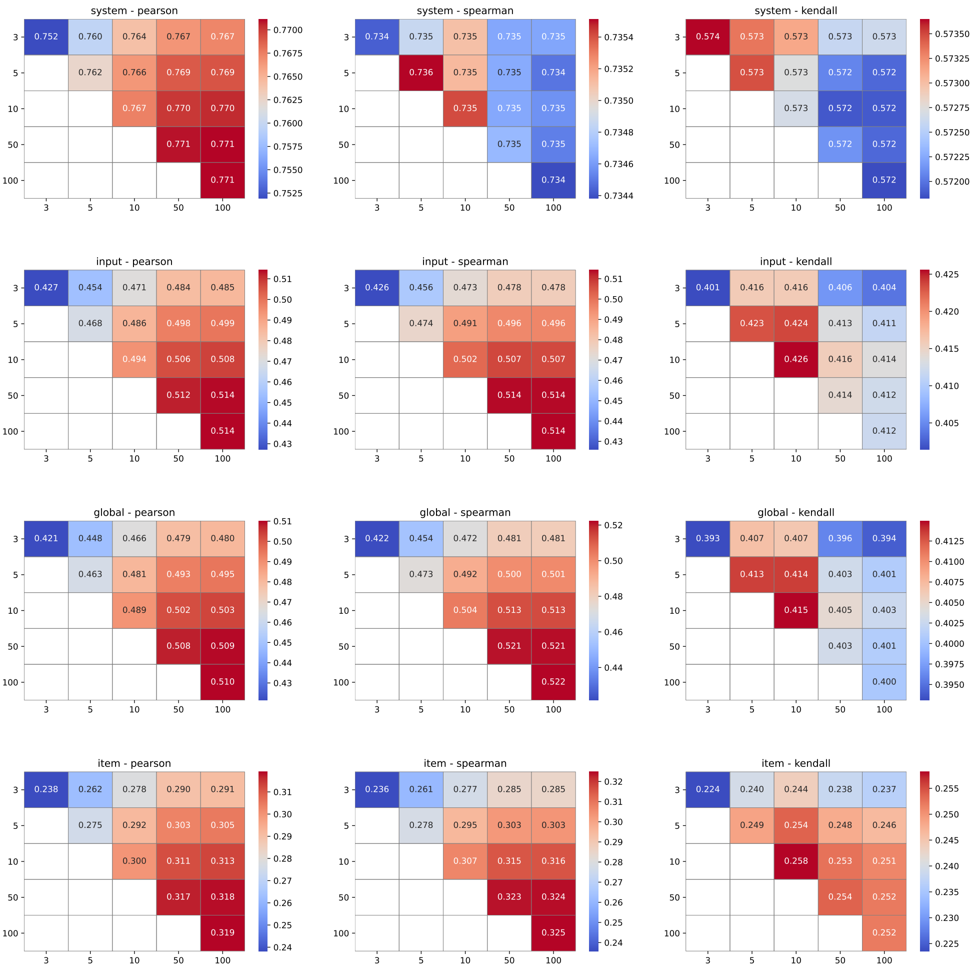

Figure 5: Simulation results of discretization when $C^m \geq C^h$ and $N = 15, M = 200$. The horizontal axis is the value of C^m and the vertical axis is the value of C^h.

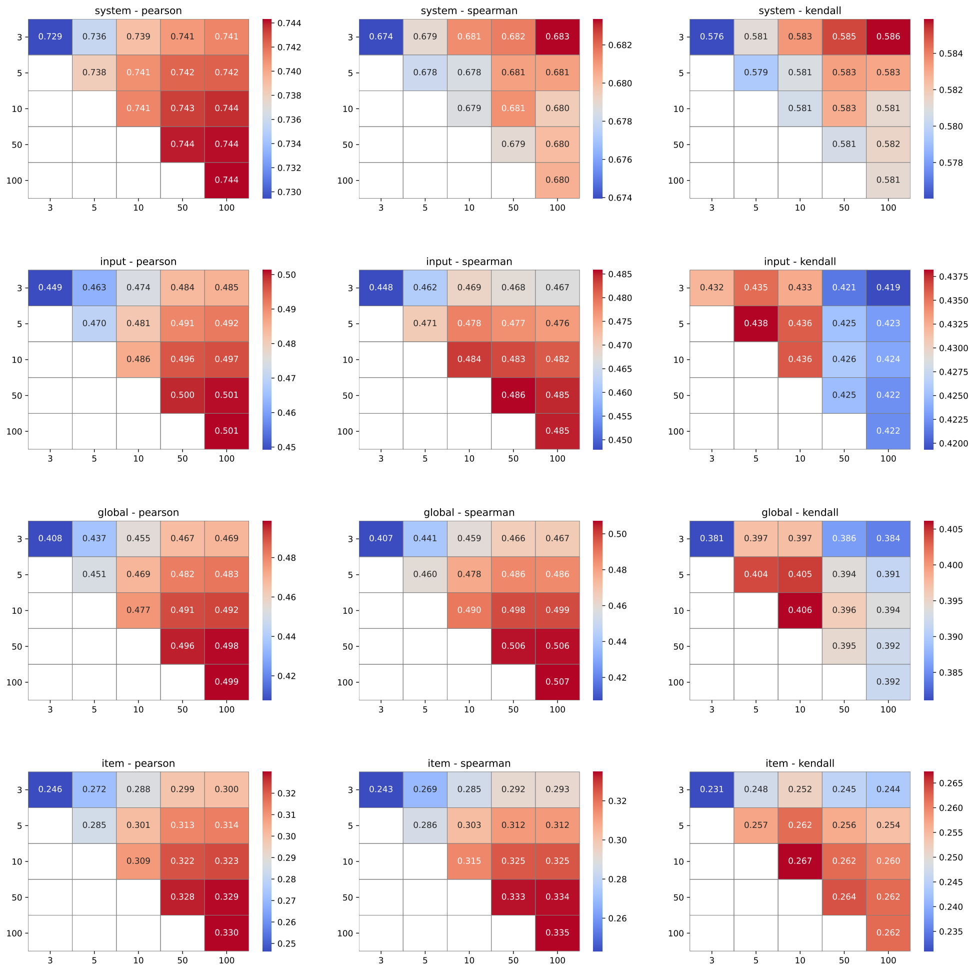

Figure 6: Simulation results of discretization when $C^m \geq C^h$ and $N = 5, M = 100$. The horizontal axis is the value of C^m and the vertical axis is the value of C^h.

Level	Function	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
System	Pearson	0.206 (4)	0.141 (4)	0.265 (5)	0.293 (6)	0.098 (4)	0.252 (4)	0.126 (4)	0.248 (4)	0.193 (4)	0.254 (4)
System	Spearman	0.219 (5)	0.181 (5)	0.194 (4)	0.257 (4)	0.164 (7)	0.314 (5)	0.226 (5)	0.365 (5)	0.288 (5)	0.337 (7)
System	Kendall	0.366 (8)	0.319 (6)	0.352 (8)	0.417 (9)	0.261 (9)	0.486 (8)	0.355 (6)	0.516 (8)	0.415 (6)	0.510 (9)
Input	Pearson	0.513 (10)	0.437 (9)	0.480 (10)	0.542 (10)	0.557 (10)	0.657 (10)	0.655 (11)	0.774 (11)	0.644 (10)	0.634 (10)
Input	Spearman	0.520 (11)	0.532 (11)	0.559 (11)	0.588 (11)	0.697 (11)	0.692 (11)	0.649 (10)	0.757 (10)	0.695 (11)	0.634 (11)
Input	Kendall	0.613 (12)	0.584 (12)	0.624 (12)	0.669 (12)	0.744 (12)	0.740 (12)	0.717 (12)	0.807 (12)	0.751 (12)	0.692 (12)
Global	Pearson	0.064 (1)	0.073 (1)	0.065 (1)	0.083 (1)	0.031 (2)	0.087 (2)	0.068 (2)	0.112 (2)	0.083 (1)	0.082 (1)
Global	Spearman	0.071 (2)	0.127 (2)	0.087 (2)	0.096 (2)	0.020 (1)	0.085 (1)	0.060 (1)	0.089 (1)	0.093 (2)	0.090 (2)
Global	Kendall	0.134 (3)	0.133 (3)	0.132 (3)	0.151 (3)	0.077 (3)	0.109 (3)	0.108 (3)	0.157 (3)	0.126 (3)	0.145 (3)
Item	Pearson	0.290 (6)	0.355 (7)	0.331 (6)	0.285 (5)	0.132 (6)	0.392 (6)	0.410 (7)	0.413 (6)	0.417 (7)	0.303 (5)
Item	Spearman	0.313 (7)	0.419 (8)	0.344 (7)	0.303 (7)	0.105 (5)	0.407 (7)	0.416 (8)	0.424 (7)	0.435 (8)	0.319 (6)
Item	Kendall	0.397 (9)	0.496 (10)	0.411 (9)	0.392 (8)	0.203 (8)	0.489 (9)	0.515 (9)	0.522 (9)	0.522 (9)	0.389 (8)

Level	Function	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
System	Pearson	0.201 (4)	0.390 (7)	0.449 (4)	0.263 (4)	0.352 (7)	0.394 (5)	0.356 (7)	0.699 (7)	0.300 (1)	0.234 (4)
System	Spearman	0.275 (5)	0.605 (8)	0.640 (8)	0.435 (7)	0.353 (8)	0.503 (8)	0.342 (6)	0.759 (8)	0.340 (2)	0.535 (8)
System	Kendall	0.425 (6)	0.674 (9)	0.740 (9)	0.474 (8)	0.417 (9)	0.627 (9)	0.396 (8)	0.878 (9)	0.403 (3)	0.715 (11)
Input	Pearson	0.749 (11)	0.765 (10)	0.857 (10)	0.731 (10)	0.754 (11)	0.792 (10)	0.776 (10)	0.882 (10)	0.909 (10)	0.610 (9)
Input	Spearman	0.747 (10)	0.785 (11)	0.862 (11)	0.733 (11)	0.743 (10)	0.801 (11)	0.794 (11)	0.882 (11)	0.922 (11)	0.665 (10)
Input	Kendall	0.796 (12)	0.816 (12)	0.885 (12)	0.772 (12)	0.781 (12)	0.832 (12)	0.832 (12)	0.895 (12)	0.934 (12)	0.721 (12)
Global	Pearson	0.126 (2)	0.076 (2)	0.150 (2)	0.120 (1)	0.153 (2)	0.138 (2)	0.087 (2)	0.228 (2)	0.534 (5)	0.079 (1)
Global	Spearman	0.100 (1)	0.073 (1)	0.149 (1)	0.134 (2)	0.082 (1)	0.108 (1)	0.080 (1)	0.217 (1)	0.467 (4)	0.093 (2)
Global	Kendall	0.149 (3)	0.128 (3)	0.189 (3)	0.208 (3)	0.155 (3)	0.177 (3)	0.184 (3)	0.273 (3)	0.563 (6)	0.168 (3)
Item	Pearson	0.532 (8)	0.258 (4)	0.518 (6)	0.412 (5)	0.278 (5)	0.402 (6)	0.317 (4)	0.539 (5)	0.848 (7)	0.316 (5)
Item	Spearman	0.528 (7)	0.264 (5)	0.500 (5)	0.418 (6)	0.266 (4)	0.358 (4)	0.333 (5)	0.531 (4)	0.866 (8)	0.345 (6)
Item	Kendall	0.614 (9)	0.346 (6)	0.596 (7)	0.513 (9)	0.345 (6)	0.460 (7)	0.463 (9)	0.603 (6)	0.903 (9)	0.420 (7)

Level	Function	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30
System	Pearson	0.236 (4)	0.228 (4)	0.254 (4)	0.300 (4)	0.254 (3)	0.148 (3)	0.105 (3)	0.286 (6)	0.171 (3)	0.228 (4)
System	Spearman	0.442 (8)	0.560 (8)	0.557 (8)	0.630 (8)	0.554 (5)	0.372 (8)	0.177 (5)	0.385 (8)	0.599 (9)	0.344 (8)
System	Kendall	0.533 (9)	0.725 (12)	0.738 (12)	0.776 (12)	0.722 (7)	0.528 (10)	0.339 (8)	0.556 (10)	0.755 (12)	0.494 (9)
Input	Pearson	0.707 (10)	0.673 (10)	0.634 (9)	0.701 (9)	0.850 (11)	0.496 (9)	0.480 (10)	0.511 (9)	0.555 (8)	0.529 (10)
Input	Spearman	0.711 (11)	0.661 (9)	0.663 (10)	0.714 (10)	0.848 (10)	0.573 (11)	0.589 (11)	0.579 (11)	0.654 (10)	0.590 (11)
Input	Kendall	0.765 (12)	0.723 (11)	0.729 (11)	0.763 (11)	0.872 (12)	0.652 (12)	0.661 (12)	0.662 (12)	0.714 (11)	0.674 (12)
Global	Pearson	0.130 (2)	0.099 (2)	0.082 (1)	0.118 (2)	0.245 (2)	0.055 (1)	0.051 (1)	0.050 (1)	0.061 (1)	0.053 (2)
Global	Spearman	0.091 (1)	0.094 (1)	0.092 (2)	0.103 (1)	0.223 (1)	0.071 (2)	0.095 (2)	0.060 (2)	0.113 (2)	0.050 (1)
Global	Kendall	0.145 (3)	0.207 (3)	0.168 (3)	0.172 (3)	0.303 (4)	0.148 (4)	0.170 (4)	0.130 (3)	0.188 (4)	0.121 (3)
Item	Pearson	0.289 (5)	0.322 (5)	0.286 (5)	0.351 (5)	0.706 (6)	0.251 (6)	0.254 (6)	0.222 (4)	0.282 (5)	0.230 (5)
Item	Spearman	0.292 (6)	0.338 (6)	0.290 (6)	0.361 (6)	0.755 (8)	0.239 (5)	0.290 (7)	0.230 (5)	0.333 (6)	0.238 (6)
Item	Kendall	0.362 (7)	0.446 (7)	0.383 (7)	0.474 (7)	0.818 (9)	0.316 (7)	0.357 (9)	0.317 (7)	0.395 (7)	0.324 (7)

Table 7: DP values of different correlation measures on all meta-evaluation datasets using William's test, the lower the better. Each column "DN" shows the result on a meta-evaluation dataset, and the mapping to the original dataset is shown in Table 5 in the appendix.

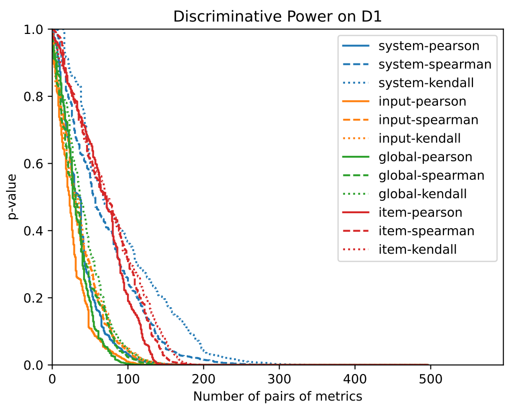

Figure 7: The p-value curves of correlation measures on meta-evaluation D1.

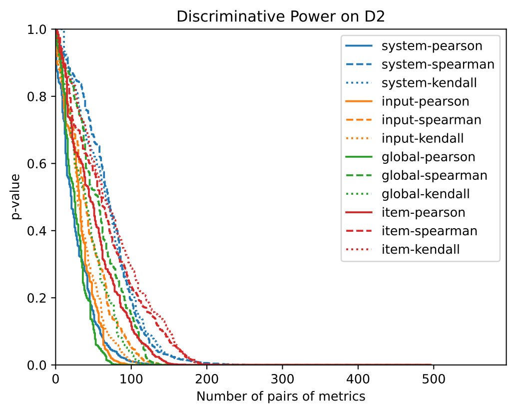

Figure 8: The p-value curves of correlation measures on meta-evaluation D2.

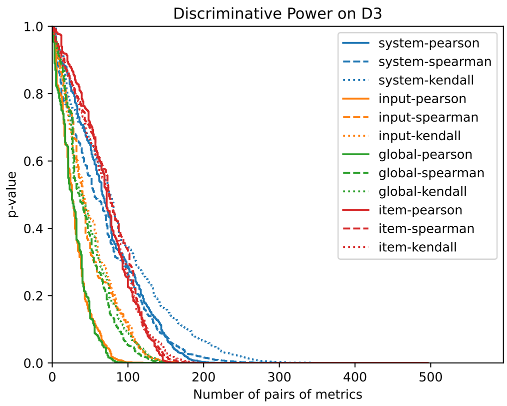

Figure 9: The p-value curves of correlation measures on meta-evaluation D3.

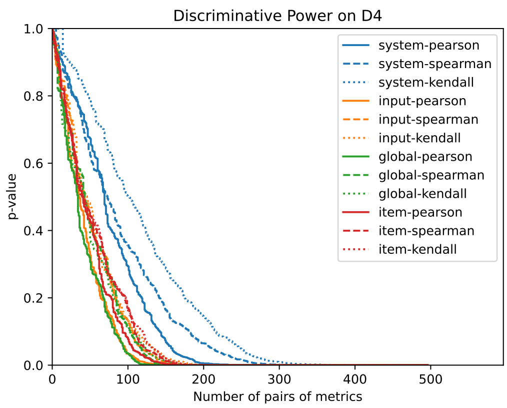

Figure 10: The p-value curves of correlation measures on meta-evaluation D4.

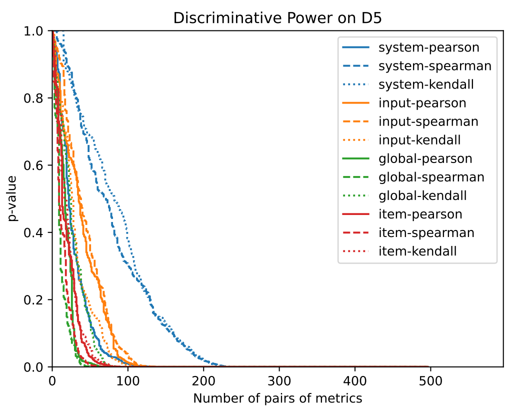

Figure 11: The p-value curves of correlation measures on meta-evaluation D5.

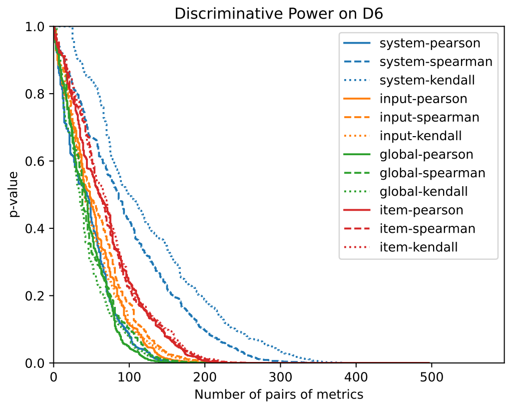

Figure 12: The p-value curves of correlation measures on meta-evaluation D6.

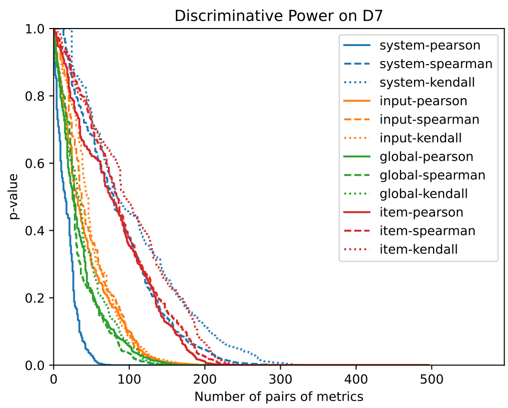

Figure 13: The p-value curves of correlation measures on meta-evaluation D7.

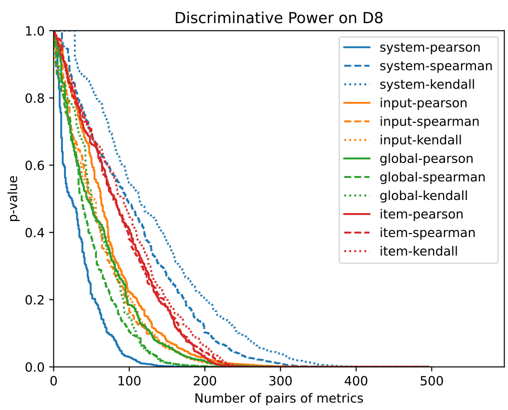

Figure 14: The p-value curves of correlation measures on meta-evaluation D8.

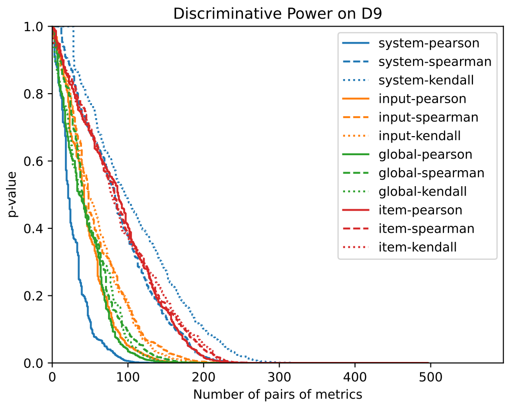

Figure 15: The p-value curves of correlation measures on meta-evaluation D9.

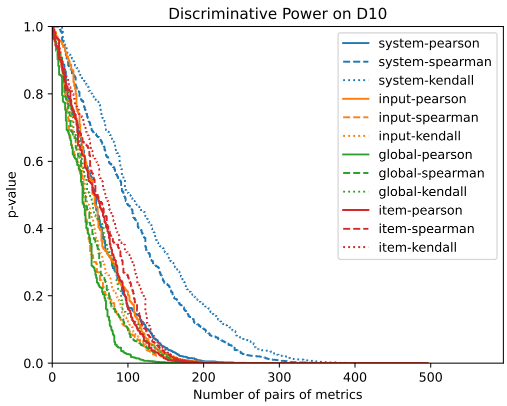

Figure 16: The p-value curves of correlation measures on meta-evaluation D10.

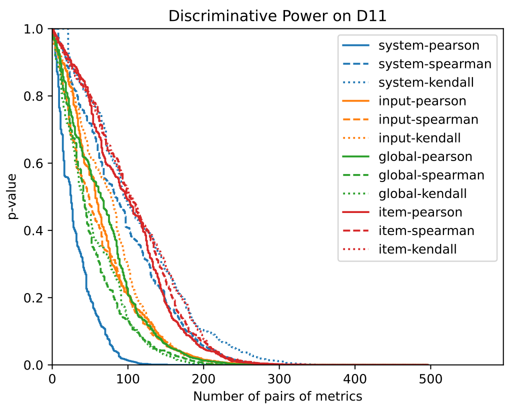

Figure 17: The p-value curves of correlation measures on meta-evaluation D11.

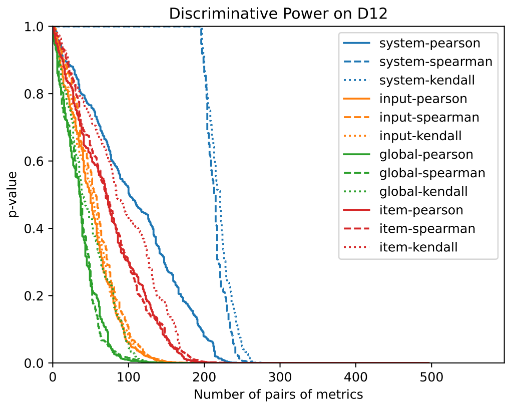

Figure 18: The p-value curves of correlation measures on meta-evaluation D12.

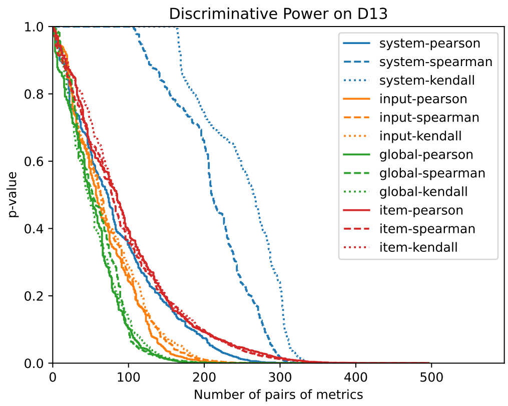

Figure 19: The p-value curves of correlation measures on meta-evaluation D13.

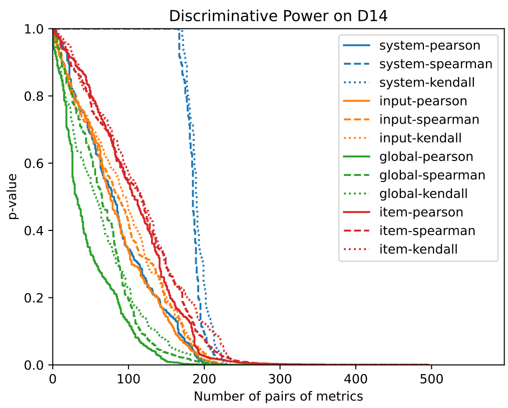

Figure 20: The p-value curves of correlation measures on meta-evaluation D14.

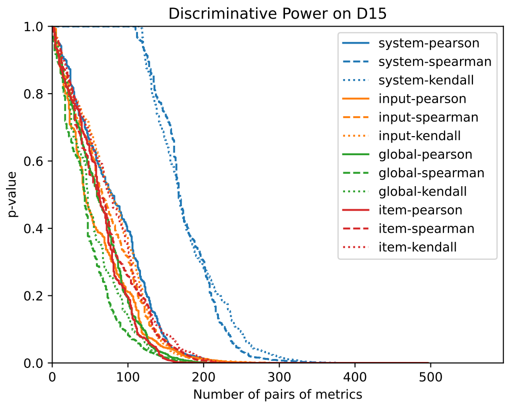

Figure 21: The p-value curves of correlation measures on meta-evaluation D15.

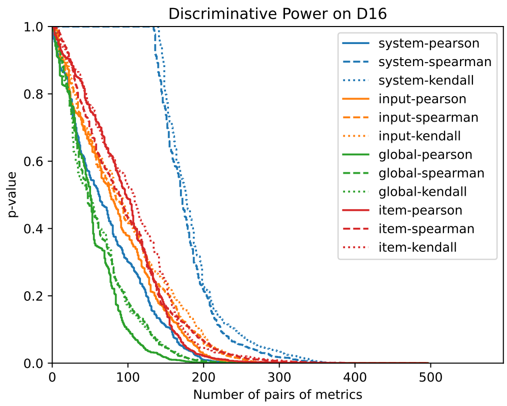

Figure 22: The p-value curves of correlation measures on meta-evaluation D16.

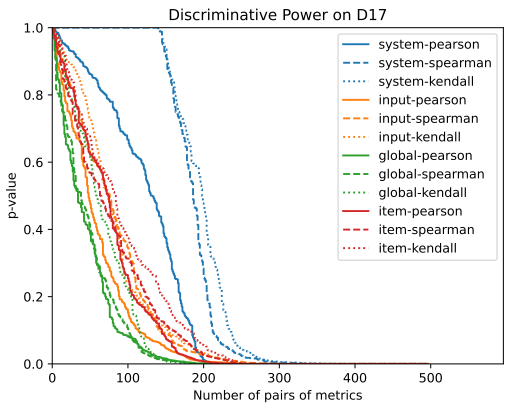

Figure 23: The p-value curves of correlation measures on meta-evaluation D17.

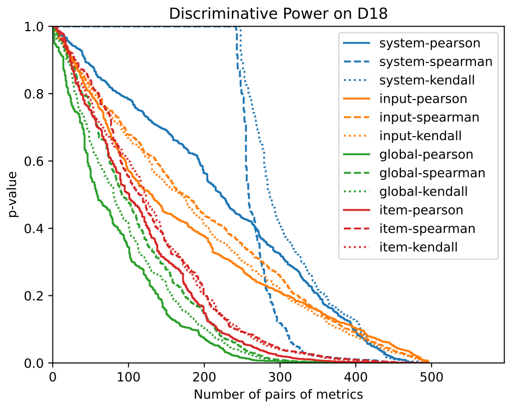

Figure 24: The p-value curves of correlation measures on meta-evaluation D18.

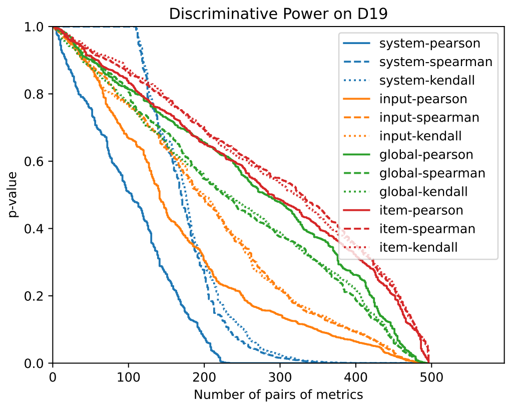

Figure 25: The p-value curves of correlation measures on meta-evaluation D19.

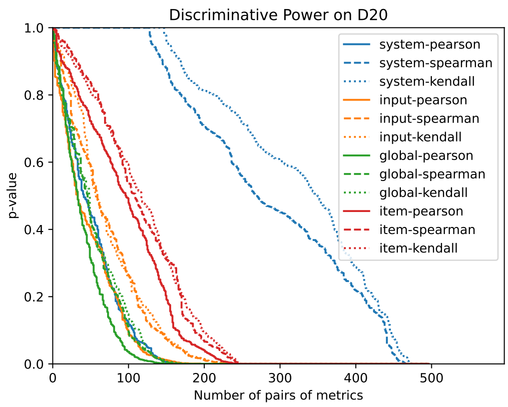

Figure 26: The p-value curves of correlation measures on meta-evaluation D20.

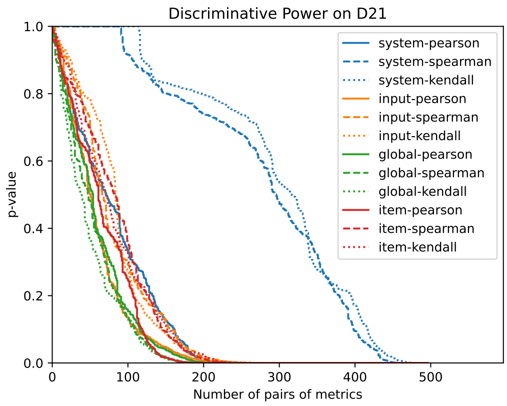

Figure 27: The p-value curves of correlation measures on meta-evaluation D21.

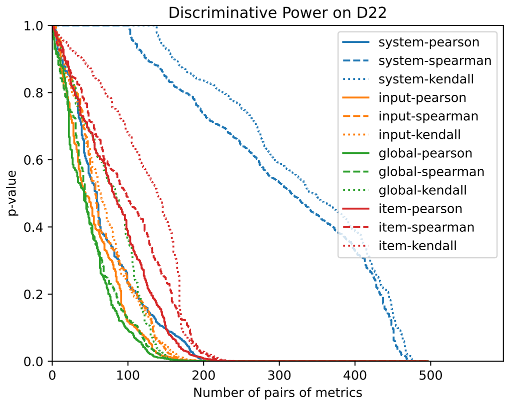

Figure 28: The p-value curves of correlation measures on meta-evaluation D22.

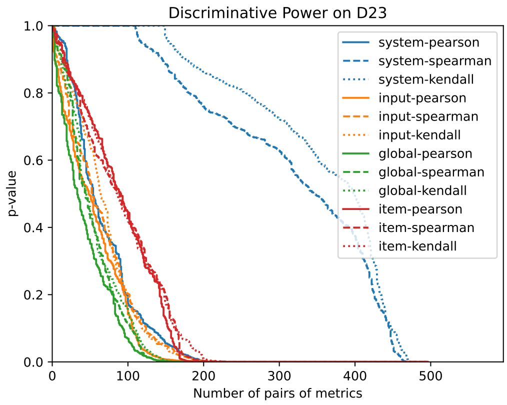

Figure 29: The p-value curves of correlation measures on meta-evaluation D23.

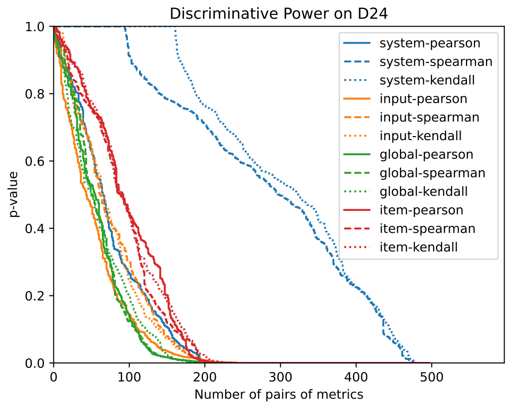

Figure 30: The p-value curves of correlation measures on meta-evaluation D24.

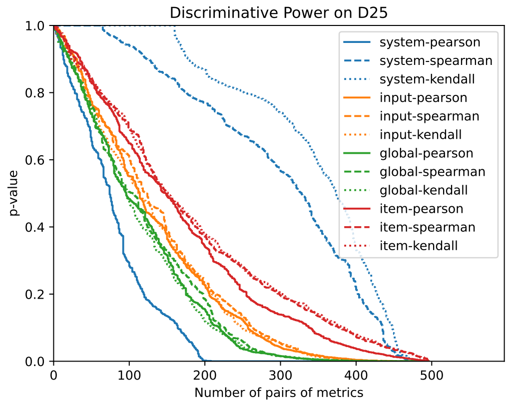

Figure 31: The p-value curves of correlation measures on meta-evaluation D25.

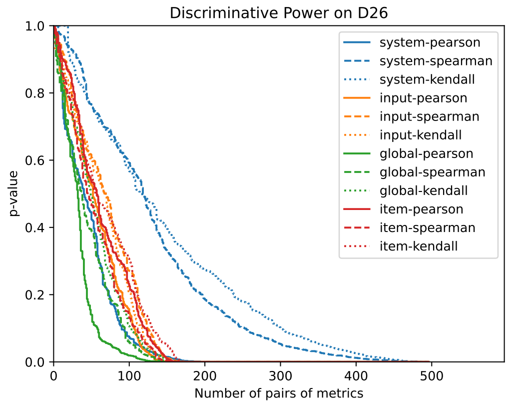

Figure 32: The p-value curves of correlation measures on meta-evaluation D26.

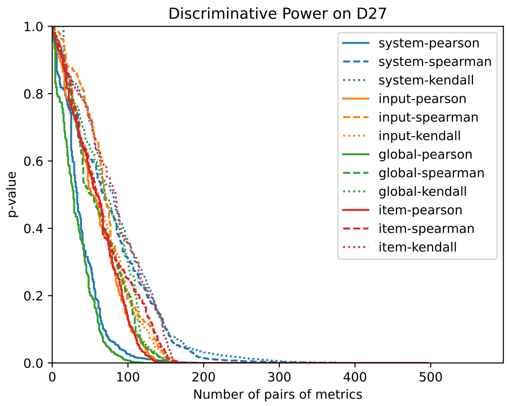

Figure 33: The p-value curves of correlation measures on meta-evaluation D27.

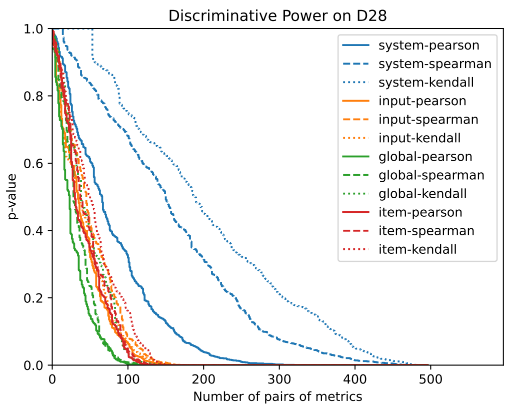

Figure 34: The p-value curves of correlation measures on meta-evaluation D28.

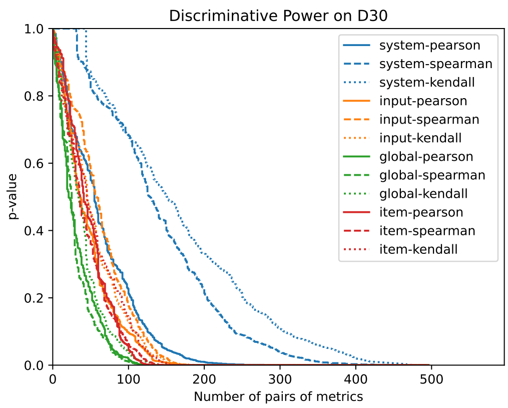

Figure 36: The p-value curves of correlation measures on meta-evaluation D30.

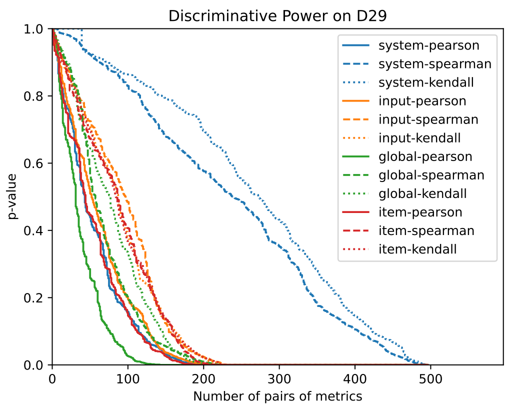

Figure 35: The p-value curves of correlation measures on meta-evaluation D29.