Confidence and Stability of Global and Pairwise Scores in NLP Evaluation

Georgii Levtsov[*]
Neapolis University Pafos / JetBrains
g.levtsov.1@nup.ac.cy

Dmitry Ustalov
JetBrains
dmitry.ustalov@jetbrains.com

Abstract

With the advent of highly capable instruction-tuned neural language models, benchmarking in natural language processing (NLP) is increasingly shifting towards pairwise comparison leaderboards, such as LMSYS Arena, from traditional global pointwise scores (e.g., GLUE, BIG-bench, SWE-bench). This paper empirically investigates the strengths and weaknesses of both global scores and pairwise comparisons to aid decision-making in selecting appropriate model evaluation strategies. Through computational experiments on synthetic and real-world datasets using standard global metrics and the popular Bradley–Terry model for pairwise comparisons, we found that while global scores provide more reliable overall rankings, they can underestimate strong models with rare, significant errors or low confidence. Conversely, pairwise comparisons are particularly effective for identifying strong contenders among models with lower global scores, especially where quality metrics are hard to define (e.g., text generation), though they require more comparisons to converge if ties are frequent. Our code and data are available at https://github.com/HSPyroblast/srw-ranking under a permissive license.

1 Introduction

Modern natural language processing (NLP) benchmarks are often represented as pairwise comparison leaderboards, as seen in projects like LMSYS Arena (Chiang et al., 2024) and AlpacaEval (Dubois et al., 2024). This trend has emerged due to the development of highly capable instruction-tuned large language models (LLMs) that output *textual* rather than categorical responses on open-ended questions. Earlier methods could be reasonably evaluated using static datasets or individual benchmarks. However, modern methods require up-to-date benchmarks that incorporate live feedback from both humans and machines (Faggioli et al., 2024). Previous benchmarks, such as GLUE (Wang et al., 2019), BIG-bench (Srivastava et al., 2023), and SWE-bench (Jimenez et al., 2024) or its live-benchmark versions, relied on global pointwise scores, prompting further research into the best approach for NLP benchmarking. But what method is most effective, and in which cases?

In this work, we empirically examine the strengths and weaknesses of pairwise comparisons and global scores. The *goal* of this study is to aid decision-making in selecting the appropriate model evaluation approach, which leads to the two following *research questions*:

RQ1. What are the strengths and limitations of global and pairwise evaluation criteria?

RQ2. Which approach is more suitable for classification problems with binary outputs and for problems where decision values (logits) or textual outputs are available?

To address these research questions, we conducted a series of computational experiments using both synthetic and realistic datasets that were distributed under permissive licenses and included model decision scores. For global evaluation scores, we selected metrics that are widely used in natural language processing and other machine learning tasks. These include accuracy, F-score, and the area under the receiver operating characteristic curve (ROC AUC) for classification tasks, as well as character-level F-score (Popović, 2015, chrF), edit distance (ED) *aka* Levenshtein distance, and word error rate (WER) for text generation tasks.

Our findings show that while global scores provide more reliable rankings of models, they tend to underestimate strong models that make rare but significant errors or have modest confidence in their

[*]The work was done during the author's internship at JetBrains.

responses. In contrast, pairwise comparisons are particularly effective for identifying strong models among those with relatively low overall scores, especially in cases where the quality metric is difficult to define—such as in text generation, which has been popularized since the release of highly-capable generative models like GPT-3 (Brown et al., 2020) and more advanced models.

The remainder of the paper is organized as follows. In Section 2, we review the related work. In Section 3, we outline the background of our study and formulate the problem. In Section 4, we describe the datasets used in our study. In Section 5, we examine the scoring stability of pairwise comparisons in the case of similar model outputs (RQ1). In Section 6, we analyze scoring stability in extreme cases of model confidence (RQ2). In Section 7, we summarize our findings and provide recommendations for using global scores and pairwise comparisons in model selection. Finally, in Section 8, we conclude with final remarks and present a flowchart to guide decision-making. Appendices A, B, and C contain supplementary information about the model scores in different settings that we tried in our work.

2 Related Work

Earlier work by Fürnkranz and Hüllermeier (2003) was focused on using pairwise comparisons (rankings) to train binary classifiers for ranking tasks, while Broomell et al. (2011) explored the use of pairwise model comparisons to identify groups of tasks where each model performs best. Maystre and Grossglauser (2017) shown that an optimal ranking of models can be achieved in a linearithmic number of comparisons, inspired by the quicksort algorithm. Nariya et al. (2023) specifically examined the use of pairwise comparisons for small datasets and studied how individual outliers and confounders impact performance estimates.

In contrast to these studies, our work aimed to identify specific scenarios in which pairwise rankings failed or behaved inconsistently, as well as cases in which they provided valuable insights across different task types, namely text classification and text generation.

3 Problem Formulation

Suppose we are given a set of models M and an evaluation dataset X, where for each element $x_i \in X$, the ground truth labels G and the model predictions $M_i(x_i)$ are known in advance. Our objective is to establish a partial order on M. As is common in NLP, this can be done using either global scores or pairwise comparisons. Examples of global scores include widely-used evaluation metrics such as accuracy, ROC AUC, and F-score, while examples of pairwise comparison methods include Bradley and Terry (1952), Elo (1978), Newman (2023), and others. We are interested in understanding the reasons behind differences in rankings produced by various methods, so we can effectively leverage the strengths of these metrics.

Global Scores. For global scores, a function $f(M_i, G) \to \mathbb{R}$, called an *evaluation score*, assigns a numerical score to each model, and the ranking is determined by a permutation P such that

$$f(M_{p_1}, G) \geq f(M_{p_2}, G) \geq \cdots \geq f(M_{p_m}, G).$$

Note that we conducted our experiments on global scores using evaluation measures implemented in scikit-learn (Pedregosa et al., 2011), edit distance and word error rate from JiWER (Morris et al., 2004), and chrF from sacreBLEU (Post, 2018) libraries for Python.

Pairwise Comparisons. For pairwise comparisons, a function $f(T) \to P$ derives a ranking from a sequence of pairwise comparisons (M_i, M_j, w), where w indicates whether M_i wins, M_j wins, or the comparison results in a tie. In our case, each test sample x_t provides $\binom{m}{2}$ pairs of models through an auxiliary function

$$g(M_i(x_t), M_j(x_t), G(x_t)) \to \{M_i, M_j, 0\},$$

and the resulting comparisons are aggregated into the global score, usually indicating the probability of each model winning against the others.

For pairwise comparisons, we used the widely known Bradley and Terry (1952) ranking model *aka* BT due to its popularity and simplicity. Although other models such as Borda count (de Borda, 1781), Elo rating (Elo, 1978), TrueSkill (Herbrich et al., 2006), and Rank Centrality (Negahban et al., 2017) are also widely used, we chose BT due it its simplicity and popularity. We intentionally did not use Elo or TrueSkill, as their outcomes depend on the order of comparisons,[1] which is more appropriate for competitive games than for time-insensitive model evaluation. Bradley and

[1] https://www.cip.org/blog/llm-judges-are-unreliable

Dataset	Response	# of examples	# of methods	# of pairs
Jigsaw (Adams et al., 2017)	Categorical	63,812	9	2,297,232
SST-5 (Socher et al., 2013)	Categorical	2,210	8	61,880
CEval (Nguyen et al., 2024)	Textual	488	6	7,320

Table 1: Descriptive statistics of the datasets used in our study; note that Jigsaw and SST-5 are classification datasets and CEval is a text generation dataset. Numbers of examples and methods are taken from the original test datasets and the corresponding baselines. The number of generated pairs is added by us.

Terry (1952) is a probabilistic model that estimates a set of latent parameters $p_1, \ldots, p_m$ such that the probability that model M_i outperforms model M_j is given by

$$P(M_i \succ M_j) = \frac{p_i}{p_i + p_j}.$$

We defined $M_i \succ M_j$ to mean that the output of i-th model is closer to the correct answer than that of the j-th model. We computed the BT scores considering each tie as a half-win and half-lose for both compared items. In our work, we used the implementation of the model from the Evalica library (Ustalov, 2025).

4 Datasets

We conducted experiments on two classification benchmarks, Jigsaw by Google (Adams et al., 2017)[2] and Stanford Sentiment Treebank (Socher et al., 2013) *aka* SST-5, and on one textual benchmark called CEval (Nguyen et al., 2024); see Table 1 for details. We selected these datasets because they provided model outputs for individual examples (including decision-function values), were widely used in the research community, and were available under permissive licenses. We used only test subsets of all datasets. In addition, we ran a series of trials on synthetic and mixed datasets combining both synthetic and real labels.

For each test instance, we compared the outputs of m different models in a pairwise fashion, yielding $\binom{m}{2}$ model pairs. For each pair, we then drew $12m\log(m)$ comparisons at random with replacement,[3] or else used all available test instances if their count was smaller. Finally, we applied these sampled comparisons to build a Bradley–Terry ranking of the models.

Jigsaw. We derived a dataset from a popular binary classification dataset for detecting text toxicity called Jigsaw (Adams et al., 2017). We collected the submission files for nine different models from the leaderboard published by their authors.[4] Since the authors did not provide ground-truth responses for the test subset of the dataset, we reconstructed them by taking the majority vote from the model-generated responses. These models included the winning method (TTA + PL), DistilBERT, JMTC-20, NB-SVM, XGBoost, XLM-R Conv1D, XLM-R, XLM-RoBERTa Bayesian, and XLM-RoBERTa. Appendix A contains scores exhibited by these models in several variations of this dataset that we created for our experiments. Although the Jigsaw suite of benchmarks contained other tasks than toxicity detection, e.g., classification bias detection,[5] we found similar results on them during prototyping. Thus, we decided not to include them in our study.

SST-5. We used the Stanford Sentiment Treebank dataset (Socher et al., 2013, SST-5),[6] a multi-class benchmark for reviews spanning five sentiment categories. To obtain model predictions, we followed the methodology of Gösgens et al. (2021) and reran eight open-source baselines.[7] These baselines included: dictionary-based methods VADER and TextBlob, traditional machine learning methods like logistic regression and support vector machine (SVM), *fast*Text classifier (Joulin et al., 2017), and deep learning classifiers: BERT and ELMo with Flair (Akbik et al., 2019) and fine-tuned BERT with

[2] https://jigsaw.google.com/

[3] We adopted the linearithmic sampling strategy of Maystre and Grossglauser (2017) and found through prototyping that a multiplier of 12 gave the best performance.

[4] https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/code?competitionId=8076&sortBy=scoreDescending&excludeNonAccessedDatasources=true

[5] https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/code?competitionId=12500&sortBy=scoreDescending&excludeNonAccessedDatasources=true

[6] https://nlp.stanford.edu/sentiment/

[7] https://github.com/prrao87/fine-grained-sentiment

Measure	Acc	AUC	BT	F_1	BT_{bin}
Acc	1.00	0.90	−0.23	0.77	0.93
AUC	0.90	1.00	0.03	0.87	0.83
BT	−0.23	0.03	1.00	0.22	−0.28
F_1	0.77	0.87	0.22	1.00	0.83
BT_{bin}	0.93	0.83	−0.28	0.83	1.00

Table 2: Spearman (1904) correlations between model scores in Jigsaw (Adams et al., 2017).

Measure	Acc	BT	F_1	BT_{bin}
Acc	1.00	0.90	0.83	0.69
BT	0.90	1.00	0.93	0.55
F_1	0.83	0.93	1.00	0.71
BT_{bin}	0.69	0.55	0.71	1.00

Table 3: Spearman (1904) correlations between model scores in SST-5 (Socher et al., 2013).

Hugging Face (Wolf et al., 2020). Appendix B contains the exhibited scores.

CEval. For a dataset featuring textual outputs evaluated by non-classification metrics, we employed the CEval benchmark for counterfactual text generation (Nguyen et al., 2024),[8] which measured models' ability to generate text that reversed the emotional tone of the original English input. In this context, we evaluated six models from the original benchmark: Crest, Crowd, GDBA, LLaMA, Llama 2, and MICE. Appendix C presents the observed scores.

5 Sensitivity to Distributions of Decision Values

Our first point of interest was focused on the sensitivity of aggregated pairwise comparisons compared to global scores (RQ1). How can we estimate the sensitivity of these evaluations? What occurs when the models exhibit similar performance?

We investigated this by running experiments on the Jigsaw dataset (binary classification) and on SST-5 (multi-class classification). We then examined the decision values of models and used the class with the highest decision value as the model's output.

Raw Decision Values. We compared the nine Jigsaw models using accuracy (Acc), ROC AUC (AUC), Bradley–Terry (BT) and F_1 scores. For SST-5, we measured F_1, accuracy and pairwise comparisons, treating the model with the higher confidence score in each pairing as the winner. Table 2 showed that the global scores (Acc, AUC, F_1) yielded consistent, highly correlated rankings, as indicated by the Spearman (1904) correlation coefficient.

On Jigsaw, we found that the anomalous BT ranking resulted from some models, such as XGBoost, outputting only decision values of 0 or 1. This caused them to win disproportionately in pairwise comparisons and thus distorted the BT ordering. We observed the same effect on SST-5: SVM rose to the top of the Bradley–Terry ranking due to its more extreme confidence scores, even though its F_1 score lagged behind Flair-BERT, Flair-ELMo, or Transformer. Therefore, **we recommend applying pairwise comparisons only to models whose decision values share a similar domain**.

Binarized Decision Values. To evaluate our recommendation, we transformed the score-based outputs from Jigsaw and SST-5 into binary values by assigning 1 to each model's most confident response and 0 to all others, i.e., by rounding each output to the nearest integer.

This transformation yielded an 88% fraction of ties on Jigsaw, which affected the rankings derived from pairwise comparisons (denoted as BT_{bin} in Table 2), but did not change any of the rankings build using global scores. On SST-5, we observed strong correlations among accuracy, F_1, and BT rankings (Table 3), and the ordering remained stable across different random samples of pairs. Unlike Jigsaw, the larger number of classes on SST-5 resulted in a moderate proportion of ties (about two-thirds of all comparisons), which in turn contributed to the stability of the pairwise rankings. From these experiments, we concluded that **pairwise comparisons were sensitive to the distributions of decision values across the compared models**.

Binary Responses. We simulated a binary classification task to examine how binary responses influenced pairwise comparisons and global scores. Three models each produced uniform random binary outputs 1,000 times using different random seeds. An ideal evaluation metric would not have favored any model. We found that accuracy, ROC AUC and F_1 each equaled 0.5, whereas aggregated **pairwise comparisons systematically favored one specific model** due to its larger number

[8] https://github.com/aix-group/CEval-Counterfactual-Generation-Benchmark

Measure	Binary AP	Penalized AP
MAE	0.38	0.86
AUC	0.90	0.94
BT	$[0.33, 0.34]$	$[0.59, 0.66]$
F_1	0.50	0.50

Table 4: Performance metrics on the adjusted decision functions in the Jigsaw dataset (Adams et al., 2017).

Measure	Binary AP
ACC	1
BT	$[0.70, 0.71]$
F_1	0.5

Table 5: Performance metrics on the adjusted decision functions in the SST-5 dataset (Socher et al., 2013).

of evaluated pairs. Spearman (1904) correlation among all global scores was 1, while the Bradley–Terry ranking exhibited a strong inverse correlation of -0.5. These results suggested that pairwise comparison methods were ill-suited for distinguishing between highly similar (or identical) models.

6 Instability with Overly Confident Models

Our second point of interest focused on the stability of pairwise comparisons given varying model confidence in the positive class (RQ2). Instead of calculating accuracy, we computed the mean absolute error (MAE) between the binary label of the target class and the model's decision value.

Binarized Decision Values. We inflated the confidence of model decision values in the Jigsaw dataset through binarization to assess its impact on model rankings. A good evaluation score should distinguish the original models from the binarized ones, ideally ranking the originals at the top and the binarized models at the bottom.

In the Jigsaw experiments, we observed that under MAE and AUC metrics, most binarized models fell in the rankings according to the average precision score (Buckley and Voorhees, 2000). However, based on F_1, the binarized models received identical scores to the originals due to the binarization performed internally inside the models. In contrast, the Bradley–Terry rankings were disrupted by the inflated model confidences (see Table 4, Binary AP). Confidence intervals for the Bradley–Terry model, here and throughout the paper, were estimated as 95% intervals by drawing 1,000 random subsamples of $12m\log(m)$ match sets for each model pair.

Measure	Penalized AP
ED	0.37
WER	0.38
chrF	0.66
BT	$[0.66, 0.70]$

Table 6: Performance metrics on the adjusted decision functions in the CEval dataset (Nguyen et al., 2024).

Although increased model confidence might challenge the evaluation in text generation tasks, in practice **it seems difficult to alter textual outputs in a way that changed pairwise rankings without also affecting other evaluation metrics**. In the CEval experiments, both WER and chrF scores remained correlated with the Bradley–Terry pairwise rankings, even after simple manipulations such as appending random strings to the outputs (see Table 7).

Penalized Decision Values. In this experiment, we artificially perturbed the model outputs in the Jigsaw and CEval datasets using the ground-truth responses to generate a heavier tail of incorrect answers and to assess how the rankings responded to such perturbations.

For the Jigsaw dataset, we binarized the decision value whenever the model made a mistake, similarly to the previous experiment; otherwise, we left the decision values unchanged. Hence, any mistake led to a model receiving worse scores, while models without errors retained their original scores. We found that under MAE and AUC, most penalized models fell to the bottom of the rankings, whereas F_1 produced results identical to those of the earlier experiment. The Bradley–Terry rankings did not correlate well with the other metrics; nevertheless,

Measure	ED	WER	chrF	BT
ED	1.00	0.94	-0.94	-0.94
WER	0.94	1.00	-1.00	-0.89
chrF	-0.94	-1.00	1.00	0.89
BT	-0.94	-0.89	0.89	1.00

Table 7: Spearman (1904) correlations between model scores in CEval (Nguyen et al., 2024). Note that some values are negative due to inverted rankings.

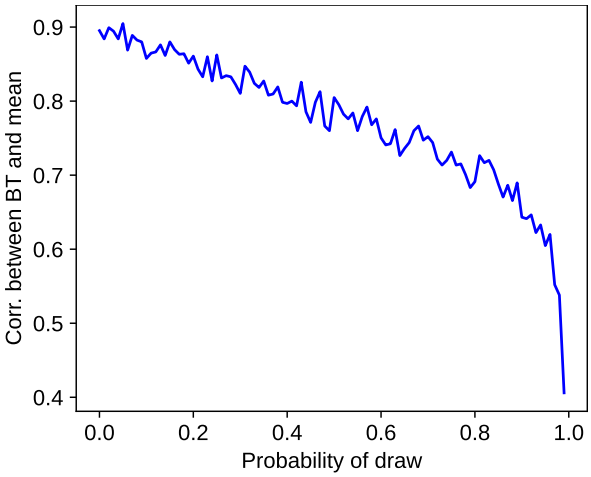

Figure 1: Dependency of the correlation between absolute and pairwise rankings in a synthetic experiment based on the CEval dataset (Nguyen et al., 2024). The results show that the Bradley–Terry model produces reliable rankings even with a large fraction of ties.

they correctly placed most original models above the penalized ones (see Table 4, Penalized AP, and a similar Table 5 for SST-5).

A similar pattern arose in the text-generation tasks. We appended random long strings to a random 5% of model outputs in the CEval dataset, which caused their distance-based global scores (ED and WER) to decline, positioning them near the bottom. However, the pairwise and chrF rankings remained largely stable (see Table 6, Penalized AP). Given that a 5% error rate can represent a substantial difference, we recommend filtering out such extreme cases or employing multiple evaluation metrics, since pairwise comparisons tend to be relatively insensitive to rare but large deviations.

From this experiment, we concluded that **pairwise comparisons can still favor promising models even when they commit rare but significant errors**.

Scored Responses. As suggested by Gösgens et al. (2021) and confirmed by our experiments, the F_1 score was a viable alternative to accuracy for binary classification tasks with an available decision function. However, ROC AUC and BT yielded more accurate results and recovered the true ranking. Nonetheless, **pairwise comparisons had to be conducted carefully to avoid favoring models that produced more confident predictions**, e.g., decision values closer to the extremes, like logits near 0 or 1.

7 Discussion

Draws in Comparisons. We noticed that Bradley and Terry (1952) rankings had performed poorly when a large fraction of comparisons resulted in draws (Section 5). They produced indistinguishable results and required a high number of observations to achieve a stable ranking, which led to high computational costs. Accuracy also tended to penalize models that made rare but significant errors. In contrast, pairwise comparisons identified such models effectively, although they sometimes demanded additional measures to ensure correctness (Section 6). Pairwise comparisons proved particularly useful for tasks which are uneasy to evaluate according to the ground-truth data, as had been confirmed by modern benchmarks (Chiang et al., 2024; Dubois et al., 2024).

In text generation tasks, ties occurred far less frequently than in classification, since evaluation metrics for generation rarely yielded identical scores. Using the CEval dataset as an example, we simulated the effect of introducing synthetic ties on the resulting rankings. More specifically, we measured the correlation between average rankings and pairwise chrF-based rankings for five models, varying the tie probability from 0 to 1 in increments of 0.01. For each probability level, we conducted 1,000 trials with $12m\log(m)$ matches per model pair. The results demonstrated that the rankings maintained a strong correlation (0.8) even when ties represented up to 50% of outcomes (see Figure 1).

However, we observed that this behavior generally depended on both the closeness of model performance and the total number of comparisons done.

Comparison Stability. To examine how the number of comparisons affects ranking stability, we constructed Bradley–Terry rankings by randomly selecting an equal number of comparisons for each pair of models, varying this number from 10 to 1000 in increments of 10. At each step, we computed the average number of changes in the ranking over 100 trials, relative to the ranking obtained using 100,000 random comparisons per pair. As mentioned earlier, we adopted the linearithmic sampling strategy proposed by Maystre and Grossglauser (2017) and settled on using $12m\log(m)$ comparisons, which provided stable results while maintaining a low computational complexity. Figure 2 presents the corresponding plot for the Jigsaw dataset, though a similar effect was observed across

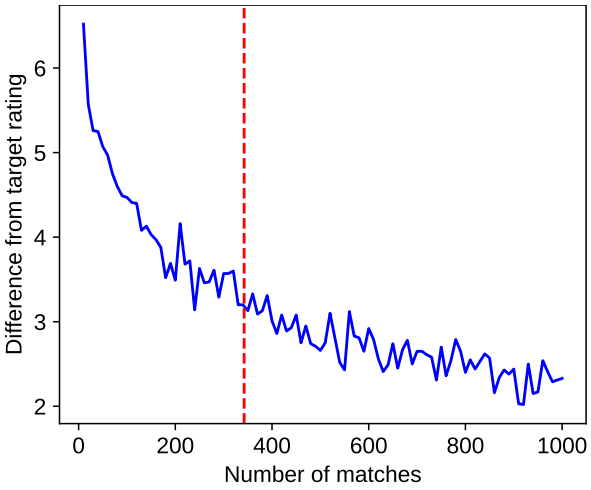

Figure 2: Comparison of stability in the Jigsaw dataset (Adams et al., 2017). The red line indicates $12m\log(m)$.

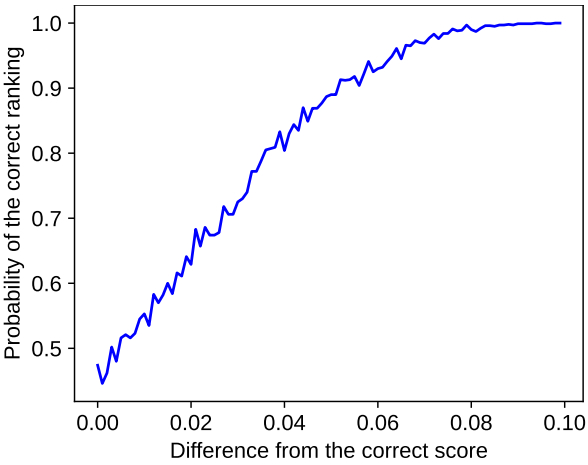

Figure 3: Dependency of probability on difference in a synthetic experiment: the larger the difference between model outputs, the better pairwise comparisons can correctly rank the models.

the other datasets as well.

Magnitude of Difference. As in the binary-response experiment described earlier, we investigated the magnitude of differences that aggregated pairwise comparisons could detect. Specifically, we examined how the probability of correct ranking depended on the difference between the decision functions of the models, such as logits or class scores. We created a grid of score differences spanning 0.9 to 1.0 in 100 steps. At each step, we subtracted the value from a randomly selected pair's scores and repeated this procedure 1,000 times. As shown in Figure 3, **pairwise comparisons perform best when the difference between model outputs is non-negligible**; for example, when there was at least a 10% difference in class probability in our synthetic example.

8 Conclusion

Our studies showed that pairwise comparisons identified potentially good models among those with poor global scores. They performed well on problems where the quality measure was difficult to define, such as text generation (RQ2). However, when a large fraction of comparisons ended in ties, the algorithm required a large number of comparisons to converge. In contrast, global scores performed better on evaluation measures that were easier to define and generally required smaller amounts of data (RQ1). Nevertheless, global scores tended to underestimate models that committed rare but significant errors. These results were consistent across synthetic datasets, multiple public datasets, and their variations.

While our study was limited to experiments on only three datasets, we believe the actionable recommendations we have discovered will advance the state of benchmarking in NLP. In addition to replicating our experiments on other datasets with different sets of models, we also find it interesting to explore which subset of the data each model performs best on, where we expect pairwise comparisons to excel. Figure 4 presents the flowchart for the model evaluation approach selection. Another possible limitation of our study was the use of well-known NLP datasets released before the wide adoption of LLMs. However, we believe that our results would generalize to newer datasets and models, as we observed the same effects consistently across all datasets, including the relatively recent textual dataset CEval. This analysis included then state-of-the-art open LLMs, such as Llama 2 and LLaMA. Running our experiments on a new multi-task dataset with frontier LLM responses would allow for a more comprehensive evaluation of the observed effects in a modern setting.

Although our experiments had been limited to three datasets, we believe that the actionable recommendations we derived could advance the state of NLP benchmarking. For future work, it would have been useful to replicate our experiments on additional datasets with diverse model sets and to examine the specific data subsets on which each model performed best, anticipating that pairwise comparisons would have excelled in those scenarios.

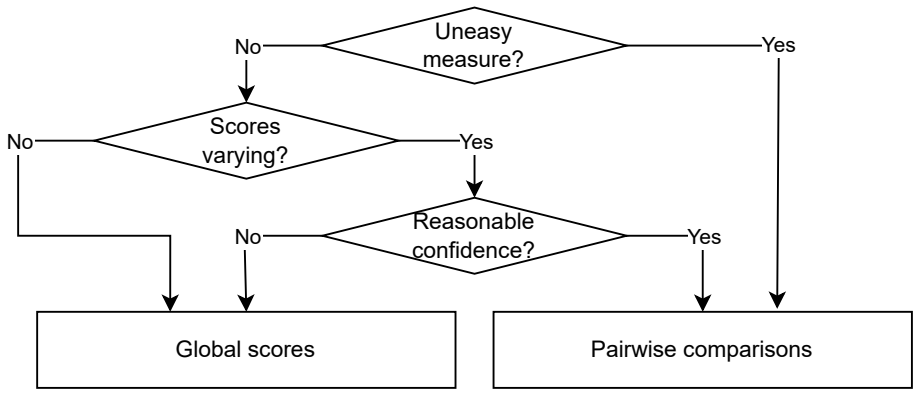

Figure 4: How to choose between global scores and pairwise comparisons? Pairwise comparisons are especially effective when the evaluation involves a difficult-to-define ("uneasy") measure, such as in text generation, or when model scores vary widely and no model shows strong confidence. In contrast, if the measure is clearly defined, the scores are relatively consistent, or some models produce more confident predictions, global evaluation metrics may be a better choice.

Acknowledgments

The authors are grateful to three anonymous reviewers whose comments allowed us to improve the manuscript. We are also grateful to the anonymous mentor who provided vital feedback during the pre-submission mentorship program at the ACL Student Research Workshop. Last but not least, we are grateful to the Internships and Academy teams at JetBrains for supporting Georgii's work.

References

CJ Adams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, Nithum Thain, and Will Cukierski. 2017. Toxic Comment Classification Challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge. Kaggle.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345.

Stephen B. Broomell, David V. Budescu, and Han-Hui Por. 2011. Pair-wise comparisons of multiple models. *Judgment and Decision Making*, 6(8):821–831.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33*, NeurIPS 2020, pages 1877–1901, Montréal, QC, Canada. Curran Associates, Inc.

Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 33–40, Athens, Greece. Association for Computing Machinery.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.

Jean-Charles de Borda. 1781. Mémoire sur les élections au scrutin. *Histoire de l'Académie royale des sciences*, pages 657–665.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-Controlled AlpacaEval: A Simple Debiasing of Automatic Evaluators. In *First Conference on Language Modeling*.

Arpad E. Elo. 1978. *The Rating Of Chess Players, Past & Present*. Arco Publishing Inc., New York.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast,

Benno Stein, and Henning Wachsmuth. 2024. Who Determines What Is Relevant? Humans or AI? Why Not Both? *Communications of the ACM*, 67(4):31–34.

Johannes Fürnkranz and Eyke Hüllermeier. 2003. Pairwise Preference Learning and Ranking. In *Machine Learning: ECML 2003*, volume 2837 of *Lecture Notes in Computer Science*, pages 145–156. Springer.

Martijn Gösgens, Anton Zhiyanov, Aleksey Tikhonov, and Liudmila Prokhorenkova. 2021. Good Classification Measures and How to Find Them. In *Advances in Neural Information Processing Systems 34*, NeurIPS 2021, pages 17136–17147, Online. Curran Associates, Inc.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems 19*, pages 569–576, Vancouver, BC, Canada. MIT Press.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Lucas Maystre and Matthias Grossglauser. 2017. Just Sort It! A Simple and Effective Approach to Active Preference Learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *ICML 2017*, pages 2344–2353, Sydney, NSW, Australia. PMLR.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768.

Maulik K. Nariya, Caitlin E. Mills, Peter K. Sorger, and Artem Sokolov. 2023. Paired evaluation of machine-learning models characterizes effects of confounders and outliers. *Patterns*, 4(8):100791.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2017. Rank Centrality: Ranking from Pairwise Comparisons. *Operations Research*, 65(1):266–287.

Mark E. J. Newman. 2023. Efficient Computation of Rankings from Pairwise Comparisons. *Journal of Machine Learning Research*, 24(238):1–25.

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2024. CEval: A benchmark for evaluating counterfactual text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 55–69, Tokyo, Japan. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Charles Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 5.

Dmitry Ustalov. 2025. Reliable, reproducible, and really fast leaderboards with evalica. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 46–53, Abu Dhabi, UAE. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR) 2019*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz,

Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Jigsaw Rankings

We present below the scores of the described models from our Jigsaw-derived dataset (Adams et al., 2017).

A.1 Raw Jigsaw Dataset (Section 5)

Model	Acc	AUC	BT	F_1	BT_{bin}
TTA + PL	0.895	0.954	0.082	0.740	0.122
JMTC-20	0.895	0.955	0.083	0.739	0.121
XLM-R	0.889	0.952	0.093	0.714	0.115
XLM-RoBERTa	0.886	0.944	0.067	0.721	0.118
XLM-R Conv1D	0.883	0.943	0.167	0.731	0.117
XLM-RoBERTa Bayesian	0.849	0.501	0.029	0.171	0.110
DistilBERT	0.835	0.882	0.144	0.523	0.105
NB-SVM	0.821	0.866	0.071	0.367	0.102
XGBoost	0.754	0.745	0.264	0.572	0.089

A.2 Binarized Jigsaw Dataset (Section 6)

Model	Accuracy	ROC AUC	BT	F_1
XGBoost	0.754	0.745	0.062	0.572
XLM-RoBERTa Bayes	0.797	0.501	0.008	0.171
NB-SVM	0.812	0.866	0.013	0.367
XLM-RoBERT	0.816	0.944	0.013	0.721
DistilBERT	0.819	0.882	0.021	0.523
XLM-R Conv1D	0.834	0.943	0.023	0.731
TTA + PL	0.846	0.954	0.015	0.740
JMTC-20	0.849	0.955	0.015	0.739
XLM-R	0.856	0.952	0.017	0.714
Binarized XGBoost	0.754	0.745	0.060	0.572
Binarized NB-SVM	0.821	0.612	0.079	0.367
Binarized DistilBERT	0.835	0.681	0.081	0.523
Binarized XLM-RoBERTa Bayes	0.849	0.499	0.089	0.171
Binarized XLM-R Conv1D	0.883	0.819	0.100	0.731
Binarized XLM-RoBERT	0.886	0.804	0.099	0.721
Binarized XLM-R	0.889	0.791	0.099	0.714
Binarized 1st Place	0.895	0.813	0.104	0.740
Binarized JMTC-20	0.895	0.811	0.101	0.739

A.3 Penalized Jigsaw Dataset (Section 6)

Model	Acc	AUC	BT	F_1
XGBoost	0.754	0.745	0.142	0.572
XLM-RoBERTa Bayesian	0.797	0.501	0.017	0.171
NB-SVM	0.812	0.866	0.040	0.367
XLM-RoBERT	0.816	0.944	0.032	0.721
DistilBERT	0.819	0.882	0.079	0.523
XLM-R Conv1D	0.834	0.943	0.088	0.731
TTA + PL	0.846	0.954	0.042	0.740
JMTC-20	0.849	0.955	0.044	0.739
XLM-R	0.856	0.952	0.053	0.714
Penalized XLM-RoBERTa Bayesian	0.751	0.502	0.013	0.171
Penalized XGBoost	0.754	0.745	0.139	0.572
Penalized XLM-RoBERT	0.773	0.625	0.026	0.721
Penalized DistilBERT	0.787	0.385	0.065	0.523
Penalized NB-SVM	0.793	0.228	0.035	0.367
Penalized XLM-R Conv1D	0.793	0.656	0.072	0.731
Penalized 1st Place	0.812	0.638	0.034	0.740
Penalized JMTC-20	0.816	0.633	0.036	0.739
Penalized XLM-R	0.827	0.594	0.045	0.714

B SST-5 Rankings

We present below the scores of the described models from the SST-5 dataset (Socher et al., 2013).

B.1 Raw SST-5 Dataset (Section 5)

Model	Acc	BT	F_1
TextBlob	0.284	0.067	0.255
VADER	0.316	0.084	0.315
Logistic Regression	0.409	0.135	0.383
SVM	0.414	0.126	0.401
*fast*Text	0.434	0.120	0.384
Flair-ELMo	0.462	0.143	0.408
Transformer	0.491	0.162	0.486
Flair-BERT	0.511	0.162	0.491

B.2 Binarized SST-5 Dataset (Section 5)

Model	Acc	BT	F_1
TextBlob	0.225	0.032	0.255
VADER	0.248	0.054	0.315
Logistic Regression	0.258	0.043	0.383
*fast*Text	0.272	0.052	0.384
Flair-ELMo	0.344	0.155	0.408
Flair-BERT	0.353	0.124	0.491
Transformer	0.360	0.154	0.486
SVM	0.384	0.386	0.401

C CEval Rankings

We present below the scores of the described models from the CEval dataset (Nguyen et al., 2024).

C.1 Raw CEval Dataset (Section 6)

Model	ED	WER	chrF	BT
Crowd	162.041	0.239	81.326	0.444
MICE	229.711	0.299	73.674	0.163
Llama 2	274.370	0.375	70.886	0.202
LLaMA	298.368	0.404	68.378	0.125
GDBA	333.184	0.540	55.427	0.017
Crest	362.584	0.477	63.324	0.049

C.2 Penalized CEval Dataset (Section 6)

Model	ED	WER	chrF	BT
Crowd	162.041	0.239	81.326	0.240
MICE	229.711	0.299	73.674	0.093
Llama 2	274.370	0.375	70.886	0.095
LLaMA	298.368	0.404	68.378	0.075
GDBA	333.184	0.540	55.427	0.025
Crest	362.584	0.477	63.324	0.023
Penalized Crowd	272.713	0.363	79.950	0.189
Penalized MICE	384.359	0.451	72.188	0.077
Penalized Llama 2	437.590	0.592	69.111	0.078
Penalized LLaMA	484.732	0.657	66.350	0.059
Penalized GDBA	475.117	0.698	54.434	0.022
Penalized Crest	458.033	0.589	62.539	0.022