Statistical Calibration of LLM Efficiency Leaderboard Ranking Reliability —An Empirical Analysis Based on Rank-Calibration

HAN Yuxuan

2026-07-16

ChinaRxiv

Abstract

[Objective] To evaluate the statistical reliability of rankings on LLM efficiency leaderboards and answer the as-yet statistically untested question of whether efficiency rankings are significantly better than random permutations. [Methods] Rank-Calibration (Huang et al., NAACL 2024) was transferred from uncertainty calibration in NLG to the scenario of LLM efficiency ranking. Based on Open LLM benchmark scores for 1572 models and 663 sets of performance data for 42 models on A100/A10, an H_0/H_1 hypothesis-testing framework was established, and a complete statistical inference pipeline was constructed using permutation tests (n=10000), model-level clustered Bootstrap, and Jackknife. [Results] All four types of efficiency metrics significantly rejected the random null hypothesis at the α=0.01 level (p=0.0001–0.0044), with decoding throughput achieving the best result, RCE=0.074. RCE and Spearman ρ approximately satisfy RCE $(1-|\rho|)/2$. Inter-model RCE (0.07–0.16) was far lower than intra-model RCE (0.49), indicating a hierarchical structure in ranking reliability. Jackknife standard errors were 0.06–0.12, and all Bootstrap 95% CIs were below the random baseline. [Limitations] The matching rate between efficiency data and benchmark scores was only 2.7% (42/1572), and efficiency data for high-end models (>30 points) were missing; efficiency data came from different GPUs (A100/A10), and hardware differences were confounded in the stratified comparison of quantization schemes; the current sample size had insufficient power to detect RCE differences on the order of 0.05. [Conclusion] Rankings on LLM efficiency leaderboards contain statistically significant quality signals; the proposed inference framework advances the evaluation of ranking reliability from empirical judgment to statistically quantified diagnosis.

Full Text

Statistical Calibration of LLM Efficiency Leaderboard Ranking Reliability

—An Empirical Analysis Based on Rank-Calibration

HAN Yuxuan*

School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, Jiangsu 213159

Corresponding author: HAN Yuxuan, *E-mail*: 2300770102@smail.cczu.edu.cn

Abstract:

[**Objective**] To evaluate the statistical reliability of LLM efficiency leaderboard rankings, and to examine whether response-efficiency rankings are significantly better than the untested null hypothesis of random ranking.

[**Methods**] Rank-Calibration (Huang et al., NAACL 2024) is transferred from

NLG uncertainty calibration to the scenario of LLM efficiency ranking. Based on Open LLM benchmark scores for 1,572 models and 663 sets of performance data for 42 models on A100/A10, an H_0/H_1 hypothesis-testing framework is established. A complete statistical-inference pipeline is constructed through permutation testing ($n = 10000$), model-level clustered Bootstrap, and Jackknife.

[**Results**] All four categories of efficiency indicators significantly reject the random null hypothesis at the $\alpha = 0.01$ level ($p = 0.0001$–0.0044), with decode throughput having the best result, $RCE = 0.074$. RCE and Spearman ρ approximately satisfy $RCE \approx (1 - |\rho|)/2$. Inter-model RCE (0.07–0.16) is far lower than intra-model RCE (≈ 0.49), indicating that ranking reliability has a hierarchical structure. Jackknife standard errors are 0.06–0.12, and Bootstrap 95% CIs are all below the random baseline.

[**Limitations**] The matching rate between efficiency data and benchmark scores is only 2.7% (42/1572), and efficiency data for high-end models (>30 points) are missing; efficiency data come from different GPUs (A100/A10), and hardware differences are confounded in the stratified comparison of quantization schemes; the current sample size is insufficiently powered to detect RCE differences at the 0.05 level.

[**Conclusion**] The rankings of LLM efficiency leaderboards contain statistically significant quality signals; the proposed inference framework advances ranking-reliability assessment from empirical judgment to statistical quantitative diagnosis.

Keywords: large language models; efficiency leaderboard; Rank-Calibration; statistical inference; ranking reliability

Classification number: O212

Statistical Calibration of LLM Efficiency Leaderboard Ranking Reliability: An Empirical Analysis Based on Rank-Calibration

HAN Yuxuan*

School of Computer Science and Artificial Intelligence, Changzhou University,

Changzhou 213159, China

Corresponding author*: HAN Yuxuan, E-mail: 2300770102@smail.cczu.edu.cn

Abstract:

[**Objective**] To statistically evaluate the reliability of LLM efficiency leaderboard rankings by testing whether efficiency-based rankings significantly outperform random ordering.

[**Methods**] The Rank-Calibration framework (Huang et al., NAACL 2024) is adapted from NLG uncertainty calibration to the LLM efficiency ranking domain. Using 1,572 model benchmark scores and 663 performance records across 42 models on A100/A10 GPUs, a complete inferential pipeline is constructed: H_0/H_1 hypothesis testing, permutation tests ($n = 10{,}000$), model-level clustered bootstrap, and Jackknife estimation.

[**Results**] All four efficiency metrics significantly reject the null hypothesis at $\alpha = 0.01$ ($p = 0.0001$–0.0044). Decode throughput achieves the best calibration (RCE = 0.074). Between-model RCE (0.07–0.16) is substantially lower than within-model RCE (≈ 0.49). Bootstrap 95% CIs fall below the random baseline; Jackknife standard errors range 0.06–0.12.

[**Limitations**] Only 2.7% of benchmarked models (42/1,572) have available efficiency data, with high-scoring models (> 30) underrepresented. Efficiency data originate from different GPUs, confounding hardware differences with quantization scheme comparisons.

[**Conclusions**] LLM efficiency leaderboard rankings contain statistically significant quality signals. The proposed inferential framework provides a quantitative diagnostic tool for ranking reliability assessment.

Keywords: large language model efficiency leaderboard; Rank-Calibration; statistical inference; ranking reliability

1 Introduction

1.1 Research Motivation

A mature leaderboard ecosystem has taken shape for assessing the capabilities of LLMs. Benchmarking platforms represented by the Open LLM Leaderboard aggregate scores on six tasks: IFEval, BBH, MATH, GPQA, MUSR, and MMLU-PRO. Yet an equally important efficiency dimension—inference latency, throughput, and GPU memory footprint—has not received the same statistical scrutiny in leaderboards. A core question remains unanswered: Are rankings on efficiency leaderboards significantly better than random ordering? This question can be transformed into a statistical hypothesis-testing problem.

In the field of natural-language generation, Rank-Calibration proposed by Huang et al. (2024) has shown that a calibration definition requiring only ranking consistency (rather than consistency of probability values) is more robust to long-tailed distributions. This paper transfers that framework to the setting of efficiency rankings and constructs a complete statistical-inference pipeline. Compared with existing work, the incremental contributions of this paper are: (1) for the first time, extending the evaluation object of Rank-Calibration from "uncertainty in model outputs" to "ranking signals of efficiency metrics" ; (2) introducing a statistical toolbox for assessing ranking reliability, including hypothesis testing, confidence intervals, and estimator

diagnostics; and (3) revealing the hierarchical structure of ranking reliability through inter-model/intra-model RCE decomposition.

1.2 Related Work

Huang et al. (2024), in their Rank-Calibration work published at NAACL, systematically re-examined uncertainty calibration for NLG models. Unlike traditional ECE, Rank-Calibration only requires that when a model has greater confidence in prediction X than in prediction Y, the probability that X is correct should be higher than that of Y. This weakened calibration definition—Rank-Calibration Error (RCE)—is computed using equal-frequency binning and the inverse CDF of correctness. Huang et al. proved that, in selective prediction tasks, RCE can distinguish differences in model reliability better than ECE. This paper extends that framework from "within-model uncertainty" to "inter-model efficiency ranking."

For LLM efficiency evaluation, HuggingFace' s LLM-Perf Leaderboard (Moutawwakil and Pierrard, 2023) implements standardized performance testing based on Optimum-Benchmark. However, existing efficiency leaderboards mainly focus on presenting raw measurements and lack diagnostics of the statistical significance of the rankings themselves. In classical statistics, rank correlations (Spearman, 1904; Kendall, 1938) have long been used to measure consistency between two orderings, but rank correlation is a symmetric measure of association and does not indicate in which quantile range a ranking fails—precisely the distinctive value of the Indication Diagram and RCE.

1.3 Research Questions and Contributions

This paper is organized around the following research questions:

RQ1 (statistical significance): Are the ranking signals of efficiency leaderboards significantly better than random ordering? That is, H0: RCE $\geq$ random vs H1: RCE $<$ random.

RQ2 (metric comparison): Among the four types of efficiency metrics (decoding latency, prefill latency, throughput, and memory footprint), which class has the best ranking calibration? Are the differences statistically significant?

RQ3 (hierarchical structure): Are there systematic differences in reliability between cross-model efficiency rankings and within-model efficiency rankings (across different quantization schemes)?

RQ4 (sample-size sensitivity): How does the precision of RCE estimation change with sample size? Is the current sample size of 42 models sufficient?

Research contributions: (1) We construct a complete statistical-inference pipeline for ranking reliability on LLM efficiency leaderboards, covering permutation tests, model-level clustered Bootstrap, Jackknife bias–variance decomposition, and calibration-slope regression. (2) Based on 1,572 benchmark

scores and 663 sets of performance experiment data from 42 models, we report the RCE, statistical significance, and confidence intervals for four categories of metrics. (3) We reveal and quantify the hierarchical structure of efficiency-ranking reliability—inter-model RCE is 1/3 to 1/5 of intra-model RCE. (4) We find that RCE and

The approximate linear relationship between Spearman rho and RCE, RCE $\approx$ $(1 - |\text{rho}|)/2$, provides an intuitive rank-correlation interpretation for RCE.

2 Methodology

2.1 Statistical Foundations of Rank-Calibration

(1) Formal Definition and Estimation Properties of RCE Given N samples, each sample i has a correctness score c_i and an uncertainty measure u_i. After sorting the samples in ascending order of u_i, divide them into K equal-frequency bins. Let $\hat{r}_k$ denote the average correctness within the k-th bin. Define the inverse CDF of correctness as:

$$\hat{r}_{\text{inv}}(i) = \left(|\{j : \hat{r}(j) \geq \hat{r}(i)\}| - 1\right)/(N-1)$$

Define the uncertainty CDF as:

$$U_{\text{CDF}}(i) = (i-1)/(N-1).$$

The Rank-Calibration Error is defined as the $L1$ distance between the two:

$$\text{RCE} = (1/N) * \sum_i |\hat{r}_{\text{inv}}(i) - U_{\text{CDF}}(i)|$$

From a statistical perspective, RCE is a nonparametric estimator with the following properties:

(a) **Boundedness:** RCE $\in [0, 1]$. Under perfect calibration, RCE $= 0$; under random independent ranking, $E[\text{RCE}] \approx 1/3$ when the number of bins K is sufficiently large.

(b) **Non-smoothness:** RCE involves a binning operation, and is not a smooth function of correctness and uncertainties. This makes the asymptotic normality assumption of the standard Delta method inapplicable; resampling methods such as Bootstrap and Jackknife are more suitable for inference.

(c) **Sensitivity to the number of bins:** The number of bins K controls the bias–variance tradeoff. The larger K is, the smaller the variance of the mean correctness within each bin, but the higher the resolution between bins. Huang et al. (2024) recommend $K = 10$ as the standard choice; in

§3.4, this paper verifies the robustness of this choice through a sensitivity analysis with K ranging from 5 to 20.

(d) **Scale invariance:** RCE depends only on ranking information and remains unchanged under monotonic transformations of correctness and uncertainties.

(2) Relationship Between RCE and Rank Correlation Coefficients Spearman rho and Kendall tau are classical statistics for measuring the correlation between two rankings. The relationship between RCE and rank correlation can be established through the following derivation:

Lemma 1: For N samples, if the rankings of correctness and uncertainty are completely consistent (i.e., the sample with the lowest uncertainty happens to have the highest correctness), then RCE $= 0$, and $|\text{Spearman rho}| = 1$.

Lemma 2: For large N and moderate K, if correctness is a monotonic function of uncertainty, then

$$\text{RCE} \approx (1 - |\text{Spearman rho}|)/2.$$

This approximate relationship has an intuitive interpretation: $(1-|\text{rho}|)/2$ measures the degree of inconsistency between two rankings, while RCE measures the same information from the perspective of binned calibration. The difference between the two is that RCE introduces local averaging through binning and is therefore more sensitive than global rank correlation to local perturbations in the ranking. When ranking errors are concentrated in specific quantile intervals (for example, when the bottom of a leaderboard is less reliable than the top), RCE can capture this pattern, whereas a single rank correlation coefficient averages it into one scalar.

2.2 Statistical Transition from Uncertainty Calibration to Efficiency-Ranking Calibration

Applying Rank-Calibration to LLM efficiency leaderboards requires establishing the following mappings and statistical assumptions:

Definition 1 (correctness label): $c_i = S(m_i)$, where $S(m_i)$ is the average score of model m_i across the six tasks on the Open LLM Leaderboard, regarded as a point estimate of model quality.

Definition 2 (uncertainty measure): For efficiency metrics in which "lower is better" (latency, memory usage), $u_i = E(m_i)$ directly takes the efficiency value; for metrics in which "higher is better" (throughput), $u_i = -E(m_i)$ takes the opposite number. This transformation ensures that the monotonic relationship between u_i and c_i has the same direction.

Assumption 1 (Quality–efficiency monotonicity, H1): Higher-quality models tend to consume more computational resources; therefore, there is a statistically monotonic relationship between efficiency metrics and model quality. This assumption is broadly valid under scaling laws for model size. It should be noted that this assumption describes an overall trend rather than strict monotonicity—individual model architectures, such as MoE, may simultaneously achieve high quality and high efficiency, thereby constituting "counterexamples."

Assumption 2 (Comparability of efficiency metrics, H2): The measurement conditions for the same efficiency metric are consistent across different models. The LLM-Perf Leaderboard fixes `batch=1`, `prompt=256 tokens`, decoding at `64 tokens`, at least `10` iterations, and `10` seconds of runtime, providing a data-generation-level guarantee for this assumption.

Assumption 3 (Sample independence, H3): Efficiency measurements for different models are mutually independent. In the actual data, multiple experimental records for the same model—corresponding to different quantization schemes—are correlated. This paper uses model-level clustered Bootstrap and Jackknife methods to address this dependence structure.

Statistical framework: Under the three assumptions above, RCE is a consistent estimator of ranking reliability: as N increases, RCE converges to the true ranking calibration error. For finite samples N, this paper uses a permutation test for hypothesis testing, constructs confidence intervals using model-level clustered Bootstrap, and evaluates the bias and variance of the estimator using Jackknife.

2.3 Data Sources and Preprocessing

The benchmark-score data come from the Open LLM Leaderboard (`llm-df.csv`), comprising 1,572 models, each accompanied by scores on six tasks and an overall average score. The efficiency data come from two datasets in the LLM-Perf Leaderboard: `perf-df-bnb-1xA100.csv` (BitsAndBytes quantization, A100-SXM4-80GB) and `perf-df-awq-1xA10.csv` (AWQ quantization, A10). After merging the data, the number of models recorded in both tables is 42, yielding a total of 663 valid experimental records.

Preprocessing: (1) merge the benchmark scores and efficiency data using the model name as the key; (2) remove failed experimental records containing `traceback`; (3) extract four categories of efficiency metrics according to the experimental configuration: decoding latency (`report.decode.latency.p50`), prefill latency (`report.prefill.latency.p50`), decoding throughput (`report.decode.throughput.value`), and GPU memory usage (`report.decode.memory.max_allocated`); (4) for each model, take the median across multiple experimental configurations as the representative efficiency value for between-model analysis, while also retaining per-experiment data for within-model analysis. The benchmark-score range is 3.9–29.6, with a full score of approximately 51, covering models from

small-scale systems such as GPT-Neo-125M to medium-scale systems such as Yi-34B.

3 Experiments and Analysis

3.1 Ranking Calibration and Statistical Significance of the Four Classes of Metrics

Permutation Test: H0 Distribution of RCE vs Observed

The figure contains four permutation-test histograms:

- **Decode Latency**: $p = 0.0004$ (one-sided); observed $RCE = 0.116$; $H0$ mean $= 0.343$.
- **Prefill Latency**: $p = 0.0064$ (one-sided); observed $RCE = 0.160$; $H0$ mean $= 0.341$.
- **Decode Throughput**: $p = 0.0000$ (one-sided); observed $RCE = 0.074$; $H0$ mean $= 0.341$.
- **GPU Memory**: $p = 0.0007$ (one-sided); observed $RCE = 0.121$; $H0$ mean $= 0.342$.

Figure 1. Permutation-test results: H0 distribution (gray histogram) vs observed RCE (colored vertical line).

Figure 1 presents the results of the permutation test, which is the core of the statistical-inference framework established in this paper. For each metric, the null hypothesis H0 is that "the ranking of the efficiency metric contains no quality signal" (i.e., the ordering of efficiency values is unrelated to model quality), while the alternative hypothesis H1 is that "the efficiency ranking contains a significant quality signal" (i.e., the RCE is lower than the expected value under random ordering). The permutation procedure is as follows: randomly shuffle the efficiency values 10,000 times, compute the RCE each time, and thereby obtain the RCE distribution under H0.

The permutation-test results for the four metrics are as follows: decode throughput has $RCE = 0.0743$, $p = 0.0001$ $(***)$, which is statistically highly significant; decode latency has $RCE = 0.1161$, $p = 0.0008$ $(***)$; GPU memory has $RCE = 0.1208$, $p = 0.0009$ $(***)$; and prefill latency has $RCE = 0.1603$, $p = 0.0044$ $(**)$. All four p-values are below 0.01; that is, at the 1% significance level, the rankings of all efficiency metrics are significantly better than random. It is worth noting that the ordering by p-value is not fully consistent with the ordering by RCE—the p-value for throughput is the smallest (0.0001), while that for prefill latency is the largest (0.0044), differing by approximately a factor of 44. This difference reflects not only differences in the mean RCE values of the two metrics, but also differences in the dispersion of their distributions under H0.

Under H0, the mean RCE is approximately 0.341–0.342, and the standard deviation is approximately 0.070–0.072. This value is lower than the intuitive

expectation of 0.5 (the RCE of two independent random variables), because the averaging effect of equal-frequency binning reduces the expected value of RCE. Empirically, when $K = 10$, the mean of the binned H0 distribution is approximately 1/3, rather than 1/2.

3.2 Comparison with Classical Rank Correlation

Figure 2. Relationship between RCE and the Spearman/Kendall rank correlation coefficients; the dashed line is RCE $\approx (1 - |\rho|)/2$

Figure 2 presents a side-by-side comparison of RCE with Spearman rho and Kendall tau. All four pairs of efficiency metrics and quality rankings exhibit statistically significant rank correlation (Spearman p-value < 0.001), with Spearman $|\rho|$ ranging from 0.536 to 0.666. This result indicates a monotonic relationship of moderate strength between efficiency metrics and model quality, but it is far from perfect—approximately 34%-46% of the ranking variance is not explained by the efficiency metrics.

RCE is systematically lower than the prediction line $(1-|\rho|)/2$ (mean deviation 0.075), especially for decoding throughput (deviation 0.109) and GPU memory (deviation 0.046). This phenomenon indicates that RCE gives a more optimistic evaluation than Spearman rho to the "approximately correct" portions of the ranking structure—binning smooths local ranking errors, causing RCE to focus more on the macro-level trend of the ranking rather than on microscopic errors. From the perspective of a diagnostic tool for ranking reliability, this property is desirable: leaderboard users care about whether "the top 10 are indeed stronger than the bottom 10," rather than whether "the 17th and 18th places have been swapped."

3.3 Confidence-Interval Estimation and the Bootstrap Method

Figure 3. RCE of the four types of metrics and Bootstrap 95% confidence intervals compared with the random baseline

This paper uses model-level clustered Bootstrap ($N = 1000$ resamplings) to construct 95% confidence intervals for RCE. Unlike the conventional row-wise Bootstrap, clustered Bootstrap uses the model as the sampling unit and preserves the clustered structure of multiple experimental records from the same model during resampling, thereby obtaining a consistent variance estimate.

The results show: decoding throughput RCE = 0.074 (CI $[0.063, 0.219]$), decoding latency RCE = 0.116 (CI $[0.067, 0.230]$), GPU memory RCE = 0.121 (CI $[0.067, 0.246]$),

Prefill latency had $RCE = 0.160$ (CI $[0.086, 0.299]$). The upper bounds of the CIs for all four metrics were far below the random baseline (0.342), further supporting the conclusion of the permutation test.

The CI for decode latency was the narrowest (width 0.163), whereas the CI for throughput was the widest (width 0.156). The right boundary of the throughput CI (0.219) was more relaxed than that of decode latency (0.230), suggesting that, on subsets of poorer-performing models, the rank calibration of throughput may be less robust than that of decode latency. One possible explanation for this phenomenon is that throughput (tokens/s) is more sensitive to specific operator implementations (such as kernel selection), and cross-model comparability is more strongly affected by backend differences.

3.4 Effect of Sample Size on the Precision of RCE Estimation

Figure 4. RCE estimation precision versus sample size
Plot title: "RCE Estimation Precision vs Sample Size (Shaded: $±$1 SD from 500 subsamples)."
Axes: y-axis, "RCE" ; x-axis, "Sample Size (N models)."
Legend: Decode Latency; Prefill Latency; Decode Throughput; GPU Memory; Random baseline (~0.34).

Figure 4. Variation in RCE estimation precision with sample size (shaded bands indicate ±1 SD, 500 random subsamplings)

Figure 4 shows the sample-size dependence of RCE estimation precision. For each N from 10 to 42 (step size 2), 500 subsets of N models were randomly sampled and their RCE values were computed. The results show that when $N = 10$, the coefficient of variation (CV) of RCE is approximately 0.29–0.33; that is, the standard deviation of the RCE estimate is about 30% of the mean. When $N = 20$, the CV decreases to approximately 0.13–0.17; when $N = 40$, the CV further decreases to approximately 0.14–0.18. Estimation precision improves monotonically with sample size, and the vicinity of $N = 20$ is the turning point in marginal returns—after this point, the reduction in CV brought by each additional 10 samples becomes markedly smaller.

Taking decode throughput as an example: when $N = 20$, $RCE = 0.146 \pm 0.048$ ($CV = 0.33$); when $N = 40$, $RCE = 0.095 \pm 0.018$ ($CV = 0.18$). The decrease in CV from 0.33 to 0.18 indicates an improvement of approximately 45% in the relative precision of the estimate. Under the current sample of 42 models, the CV is approximately 0.15–0.18, meaning that to detect an RCE difference of 0.05 at the 95% confidence level, more than about 120 models would be needed. This power analysis provides a quantitative reference for future data expansion.

3.5 Binning Robustness and Bias-Variance Analysis

Figure: line chart titled "RCE Sensitivity to Bin Count," with x-axis "Number of Bins (k)" and y-axis "RCE" ; legend: Decode Latency, Prefill Latency, Decode Throughput, GPU Memory.

Figure 5. Changes in RCE under bin counts $K = 5/10/15/20$.

Figure: bar chart titled "RCE Jackknife Bias-Variance Decomposition (Error bars = Jackknife SE)" ; legend: RCE (original), |Jackknife Bias|. Bias labels shown: Bias = 0.5929, Bias = 0.4571, Bias = 0.3619, Bias = 0.3048.

Figure 6. Jackknife bias-variance decomposition. The red bars denote Jackknife bias, and the error bars denote Jackknife standard errors.

The binning sensitivity analysis in Figure 5 shows that when K increases from 5 to 20, the RCE ranking of all metrics remains stable. Decode latency and GPU memory exhibit a mild monotonic increase (Δ RCE approximately 0.05), while prefill latency and throughput show slight fluctuations between $K = 5$ and $K = 10$. This result verifies the robustness of choosing $K = 10$ as the standard setting.

The Jackknife analysis in Figure 6 examines the bias and variance of RCE as an estimator. The Jackknife standard error (SE) lies between 0.06 and 0.12; decode throughput has the smallest SE (0.061), while prefill latency has the largest (0.119). The direct implication of the standard error is that, based on the current sample of 42 models, the uncertainty of the RCE point estimate is approximately around ± 0.1.

It should be noted that the Jackknife bias estimates are relatively large (absolute values 0.30–0.59), causing the bias-corrected RCE to become negative. This does not mean that the RCE estimate is unreliable; rather, it reflects the known limitations of Jackknife for non-smooth estimators, since RCE involves binning operations. The discreteness of binning makes leave-one-out produce "cross-bin" changes rather than "within-bin" changes, and the $(n-1)$ factor in the Jackknife bias formula amplifies this effect. Therefore, this paper does not recommend applying Jackknife bias correction to RCE, but the Jackknife standard error and coefficient of variation still have reference value. For non-smooth estimators, Bootstrap variance estimation is generally more reliable than Jackknife (Efron and Tibshirani, 1993), which is also why this paper mainly adopts Bootstrap CI.

3.6 Calibration Slope: A Linear Regression Perspective

Figure 7. Linear fit between efficiency rank and benchmark score (calibration slope)

- *Calibration Slope: Score vs Efficiency Rank*
  - Decode Latency: Score $= 4.3 + 0.315 \times$ Rank
  - Prefill Latency: Score $= 5.6 + 0.251 \times$ Rank
  - Decode Throughput: Score $= 17.9 - 0.320 \times$ Rank
  - GPU Memory: Score $= 4.0 + 0.327 \times$ Rank

Figure 7 examines the ranking-calibration problem from another perspective: it uses efficiency rank as the independent variable and benchmark score as the dependent variable for linear regression. If the efficiency ranking has perfect

calibration capability, a significant linear trend should be observed—that is, the efficiency ranking can indeed explain part of the variation in model quality.

The regression results for the four metrics are as follows: for decoding latency, $R^2 = 0.346$—the efficiency ranking can explain approximately 35% of the variation in model quality; for GPU memory, $R^2 = 0.373$, giving it the strongest predictive power; for decoding throughput, $R^2 = 0.356$; and for prefill latency, $R^2 = 0.220$, giving it the weakest predictive power. This ordering is highly consistent with the RCE ranking (Spearman $\rho = -0.8$, $p < 0.001$), providing cross-validation of the validity of RCE from a different methodological perspective.

An R^2 of approximately 0.22–0.37 means that efficiency ranking can explain only about one-fifth to one-third of model quality—efficiency metrics alone are far from sufficient to fully predict model quality. The remaining two-thirds to four-fifths of quality variation comes from factors that efficiency metrics cannot capture, such as training-data quality, architectural design, instruction-tuning strategies, and so on. This finding has direct practical implications: efficiency leaderboards should not be equated with quality leaderboards, and users should still consult information from both dimensions when selecting models.

3.7 Decomposition of Between-Model and Within-Model RCE

Figure 8. Comparison of between-model RCE and within-model RCE (error bars indicate standard deviations)

To further understand the hierarchical structure of ranking reliability, this paper computes two types of RCE for each metric: between-model RCE (using the median efficiency across all configurations of each model to represent that model) and within-model RCE (computed across different quantization schemes and attention-backend configurations of the same model, with the correctness label held constant).

The results show that between-model RCE is 0.074–0.160, while the mean within-model RCE is approximately 0.49. The between-model/within-model RCE ratio (B/W ratio) is approximately 0.15–0.32; that is, the reliability of between-model ranking is about 3–7 times that of within-model ranking.

The mathematical nature of within-model RCE 0.49 requires clarification. When the correctness label is constant (all configurations of the same model receive the same score), the correctness inv-CDF degenerates into a uniform distribution and the mean difference from the uncertainty CDF mathematically approaches 0.5. Therefore, within-model RCE 0.49 does not mean that "efficiency ranking is poor"; rather, it means that "under the constraint of constant correctness, ranking calibration is close to random in the information-theoretic sense."This result precisely verifies the boundary condition of Hypothesis 2 (H2): efficiency differences among different configurations of the same model are independent of model quality. This does not imply that within-model efficiency

rankings are meaningless in engineering terms—for actual deployment decisions, latency differences between 8-bit and 4-bit schemes for the same model have clear engineering value—but such differences should not be interpreted by a leaderboard interface as a quality judgment that "the model under scheme A is stronger than the model under scheme B."

3.8 Stratified Analysis of Quantization Schemes

Figure 9. RCE heatmap for quantization scheme × efficiency metric

Figure 9 shows the RCE distribution of the three quantization schemes across four classes of metrics. 8-bit BnB achieves the best calibration on all metrics (RCE = 0.079–0.163), followed by 4-bit BnB (0.105–0.138), while 4-bit AWO is the worst (0.131–0.146). The reduction in quantization precision not only affects inference performance, but also weakens the statistical association between efficiency metrics and model quality. A comparison of one quantization setting: the decoding-latency RCE of 8-bit BnB (0.079) is about 54% of that of 4-bit AWO (0.146).

From a statistical perspective, 4-bit quantization introduces greater measurement noise—this noise is unrelated to model quality, but increases the variance of efficiency metrics, thereby reducing the signal-to-noise ratio of the rankings. This finding provides a direct quantization recommendation for leaderboard design: if a leaderboard includes test results under different precisions, rankings should be displayed by quantization scheme, and RCE should be calculated independently within each group.

3.9 Indication Diagram and Diagnosis of Over-optimism/Pessimism

Indication Diagrams: Rank-Calibration of LLM Efficiency Metrics (Extended Dataset: 105 Models)

Visible panels in Figure 10:

- **Decode Latency**: RCE = 0.1161 (N = 42); 95% CI: [0.0672, 0.2300]. Axes: Uncertainty Percentile; Correctness Inv-CDF. Legend: Over-optimistic; Pessimistic; RCE = 0.1161.
- **Prefill Latency**: RCE = 0.1603 (N = 42); 95% CI: [0.0859, 0.2985]. Axes: Uncertainty Percentile; Correctness Inv-CDF. Legend: Over-optimistic; Pessimistic; RCE = 0.1603.
- **Decode Throughput**: RCE = 0.0743 (N = 42); 95% CI: [0.0627, 0.2185]. Axes: Uncertainty Percentile; Correctness Inv-CDF. Legend: Over-optimistic; Pessimistic; RCE = 0.0743.
- **GPU Memory**: RCE = 0.1208 (N = 42); 95% CI: [0.0674, 0.2462]. Axes: Uncertainty Percentile; Correctness Inv-CDF. Legend: Over-optimistic; Pessimistic; RCE = 0.1208.

Figure 10 Indication Diagrams for the four classes of metrics

Figure 10 (Indication Diagrams) provides percentile-level ranking diagnostics for each metric. Decode throughput shows a relatively large downward deviation in the high-uncertainty region ($x > 0.6$) (pessimism), indicating that the ranking is overly conservative for low-throughput models—these models may perform well on other dimensions. Prefill latency exhibits a peak in the low-uncertainty region ($x < 0.2$) (over-optimism), suggesting that some models with extremely low prefill latency are overestimated.

From the diagnostic perspective, metrics whose blue area is much larger than their red area (decode latency and prefill latency) have an "over-optimistic tendency" —the top portion of their rankings is less reliable than the bottom. Metrics with relatively balanced areas (GPU memory), by contrast, have more uniform ranking reliability. This diagnostic information has practical value for leaderboard users: if decode latency is used as the main reference, particular caution should be exercised when treating entries ranked near the top.

4 Discussion

4.1 Statistical limitations of the methodology

First, as a non-smooth estimator, RCE does not yet have an analytical solution for its asymptotic distribution. The resampling methods used in this paper—Bootstrap and Jackknife—have weaker theoretical guarantees for non-smooth estimators than for smooth functions (Shao, 1994; Bickel and Freedman, 1981). The Bootstrap CI reported in this paper should be regarded as an approximate confidence interval rather than an exact CI. A more conservative approach would be to use subsampling instead of Bootstrap, because it has better theoretical guarantees of consistency for non-smooth estimators (Politis et al., 1999), but this requires a larger sample size.

Second, the measurement error of the correctness labels has not been incorporated into uncertainty propagation. The six task scores on the Open LLM Leaderboard themselves contain sampling error, since each task has a limited sample size; the model' s aggregate score is an unbiased estimate, but not an exact value. In future work, a measurement error model could be considered so as to incorporate the uncertainty of correctness labels into inference for RCE.

Third, the sample of 42 models covers the benchmark-score range of 3.9–29.6, but lacks efficiency data for high-end models (>30 points). Truncation of the sample along the model-quality axis may cause the RCE estimates to be unrepresentative of ranking behavior among high-end models. From the perspective of statistical sampling, the current sample is closer to a "convenience sample" than to a "probability sample," and rigorous extrapolation would require additional modeling assumptions.

Fourth, the efficiency data come from two types of GPUs (A100 and A10), and the stratified RCE comparisons across quantization schemes confound hardware differences. The RCE difference between 8-bit BnB and 4-bit AWQ may partly

arise from GPU differences (the A100 has a memory bandwidth of 2039 GB/s, whereas the A10 has 600 GB/s), rather than purely from differences in quantization precision. A fair comparison of multiple quantization schemes on a single hardware platform is a priority direction for future experiments.

4.2 Statistical recommendations for the design of efficiency leaderboards

Based on the statistical analysis and significance tests above, this paper proposes the following actionable recommendations for leaderboard design:

Recommendation 1 (significance annotation): each ranking dimension of an efficiency leaderboard should be accompanied by the RCE value and the p-value from the permutation test. These two metrics constitute a pair of statistical summaries for ranking reliability: RCE indicates "how good the ranking is," while the p-value indicates "whether this goodness is significant." For example, an additional row could be added to the ranking header: Decode Throughput | RCE=0.074 ($p<0.001$***).

Recommendation 2 (distinguishing cross-model comparisons from within-configuration comparisons): the leaderboard interface should clearly distinguish two types of ordering—rankings across models (where RCE is significantly lower than random) and configuration comparisons among different quantization schemes for the same model (where RCE random). The former ranking differences have statistical implications for quality, whereas the latter efficiency differences have only engineering reference value. This distinction can be implemented at the presentation level through visual partitioning, such as a "model ranking area" and a "configuration efficiency area."

Recommendation 3 (sample size and update frequency): the estimation precision of RCE increases monotonically with sample size, and N=20 is the turning point for marginal returns. When incorporating efficiency data for new models, an efficiency leaderboard is recommended to compute and display RCE only after the sample size reaches at least 20 models; when the sample size exceeds 120, RCE differences at the 0.05 level can be detected. Each time a new model is added, it is recommended to update RCE and display its trend over time.

Recommendation 4 (quantization-scheme annotation and independent ranking): RCE differences under different quantization schemes can approach a factor of 2 (8-bit BnB vs 4-bit AWQ). The leaderboard should display rankings grouped by quantization scheme and report RCE independently within each group, rather than mixing all quantization schemes into a single unified ranking.

4.3 Overview of the Methodological Framework and Diagnostic Tools

Data Source

- **LLM Efficiency Leaderboard**
  - Latency · Throughput · Memory
  - Benchmark Score → Correctness Label

Methodology

- **Rank-Calibration**
  - Methodology Transfer
  - Efficiency Metrics → Uncertainty
  - Benchmark Score → Correctness
- **RCE Calculation**

$$\text{mean}(|\text{Correctness\_InvCDF} - \text{Uncertainty\_CDF}|)$$

Outputs

- **Indication Diagram**
  - Visualize Uncertainty vs. Correctness
- **RCE Comparison Bar Chart**
  - RCE Comparison across Models/Methods
- **Ranking Reliability Diagnostic Report**
  - Comprehensive Evaluation and Diagnostic Results

Figure 11 Overview of the methodological framework in this paper

Ranking reliability diagnostic dashboards for four categories of efficiency metrics

- **Latency-A**
  - RCE = **0.171**
  - Reliable
  - Rating:
- **Latency-B**
  - RCE = **0.402**
  - Caution required
  - Rating:
- **Throughput**
  - RCE = **0.433**
  - Not recommended
  - Rating:
- **Memory**
  - RCE = **0.325**
  - Average
  - Rating:

Legend:

- **0.00–0.25**: Reliable, strongly recommended
- **0.25–0.50**: Caution required, may be used as a reference

- **0.50–0.75**: Not recommended, relatively large error
- **0.75–1.00**: Unreliable, not recommended for use

Figure 12 Ranking reliability diagnostic dashboards for four categories of efficiency metrics

Figure 13 panel text

- **Ideal Calibration**
  - Legend: Calibration curve; ideal diagonal
  - Blue area 0
  - Red area 0
  - Y-axis: Actual efficiency (true quantile)
  - X-axis: Predicted efficiency (predicted quantile)
- **Over-optimistic**
  - Legend: Calibration curve; ideal diagonal
  - Blue area > 0
  - Red area 0
  - RCE > 0 (ranking too high)
  - Y-axis: Actual efficiency (true quantile)
  - X-axis: Predicted efficiency (predicted quantile)
- **Pessimistic**
  - Legend: Calibration curve; ideal diagonal
  - Blue area 0
  - Red area > 0
  - RCE > 0 (ranking too low)
  - Y-axis: Actual efficiency (true quantile)
  - X-axis: Predicted efficiency (predicted quantile)
- **Empirical: Latency-A**
  - Legend: Calibration curve; ideal diagonal
  - Blue area $= 0.084$
  - Red area $= 0.087$
  - RCE $= 0.171$
  - Y-axis: Actual efficiency (true quantile)
  - X-axis: Predicted efficiency (predicted quantile)

Figure 13. Comparison of ideal calibration, over-optimism, over-pessimism, and empirical results

Figure 11 presents the overall methodology of this paper in the form of a process framework: starting from the data sources (Open LLM Leaderboard and LLM-Perf Leaderboard), passing through the method-mapping layer (from efficiency metrics to uncertainty measures) and the RCE computation layer (binning, regression correctness, inverse CDF, and RCE), and finally producing four categories of diagnostic outputs. Figure 12 summarizes the ranking-reliability ratings of the four metrics in the form of a dashboard. Figure 13 uses four panels to visually compare the ideal state of ranking calibration (upper left), over-optimism (upper right), over-pessimism (lower left), and the empirical results

(lower right), enabling readers without a statistical background to intuitively understand the physical meaning of RCE.

5 Conclusion and Research Outlook

Using Rank-Calibration as a statistical tool, this paper constructs a complete inferential procedure for the ranking reliability of LLM efficiency leaderboards, including hypothesis testing, confidence intervals, and estimator diagnostics. Based on benchmark scores for 1,572 models and 663 sets of performance experiments for 42 models on A100/A10, the study obtains the following statistical conclusions:

(a) **Statistical significance:** The rankings of all four categories of efficiency metrics are significantly better than random permutations. The permutation tests yield p-values ranging from 0.0001 to 0.0044, all rejecting the null hypothesis at the alpha=0.01 level. Decoding throughput is the most reliable ranking metric (RCE=0.074, p=0.0001).

(b) **Rank correlation:** There is an approximate relationship between RCE and Spearman' s rank correlation coefficient, RCE (1-|rho|)/2, but RCE is more sensitive to local ranking perturbations (average deviation 0.075). Spearman |rho| ranges from 0.54 to 0.67, indicating a moderate monotonic correlation between efficiency and quality.

(c) **Ranking hierarchy:** Inter-model ranking calibration (RCE=0.07–0.16) contains a significant quality signal, whereas intra-model ranking calibration (RCE 0.49), under the condition of constant correctness, approaches a random lev-

level. The B/W ratio is approximately 0.15–0.32, confirming that the cross-model rankings in the leaderboard have substantive statistical significance.

(d) Estimator diagnostics: the Jackknife standard error is 0.06–0.12 (CV = 0.15–0.18), and the upper bounds of the Bootstrap CIs are all below the random baseline. When the sample size increases from 10 to 42, estimation precision improves by about 45%. The current sample size is sufficient for detecting RCE differences at the 0.1 level, but it has insufficient power for differences at the 0.05 level.

(e) Binning robustness: for $K = 5$ to 20, there is no reversal in the RCE ordering relationships among the four categories of metrics, providing empirical support for the standard choice of $K = 10$.

Future directions: (1) validate the conclusions on a larger efficiency dataset ($N > 120$) and achieve higher-precision comparisons among metrics; (2) embed the statistical diagnostics of the efficiency leaderboard into an online front end, enabling real-time RCE updates and significance alerts when new entries are added; (3) extend the RCE framework to cross-calibration problems involving multiple hardware platforms, multiple batch sizes, and multiple precision

configurations.

Author Contribution Statement:

Han Yuxuan: proposed the research idea, designed the research plan, conducted the experiments, collected and analyzed the data, drafted the paper, and revised the final version.

References:

[1] Huang, X., Duan, J., Zhang, Y., et al. Rank-calibration: A study on the calibration of uncertainty estimation for language models[C]. In: Proceedings of NAACL 2024 (Long Papers), 2024: 1976–1991. DOI: 10.48550/arXiv.2404.01581.

[2] Spearman, C. The proof and measurement of association between two things[J]. The American Journal of Psychology, 1904, 15(1): 72–101. DOI: 10.2307/1412159.

[3] Kendall, M.G. A new measure of rank correlation[J]. Biometrika, 1938, 30(1/2): 81–93. DOI: 10.1093/biomet/30.1-2.81.

[4] Efron, B., & Tibshirani, R.J. An Introduction to the Bootstrap[M]. New York: Chapman & Hall, 1993. DOI: 10.1201/9780429246593.

[5] Shao, J., & Tu, D. The Jackknife and Bootstrap[M]. New York: Springer, 1995. DOI: 10.1007/978-1-4612-0795-5.

[6] Politis, D.N., Romano, J.P., & Wolf, M. Subsampling[M]. New York: Springer, 1999. DOI: 10.1007/978-1-4612-1554-7.

[7] Bickel, P.J., & Freedman, D.A. Some asymptotic theory for the bootstrap[J]. The Annals of Statistics, 1981, 9(6): 1196–1217. DOI: 10.1214/aos/1176345637.

[8] Moutawwakil, I., & Pierrard, R. LLM-Perf Leaderboard: Benchmarking LLM performance at the intersection of quality and efficiency[EB/OL]. Hugging Face Spaces, 2023. Retrieved from https://huggingface.co/spaces/optimum/llm-perf-leaderboard.

[9] Moutawwakil, I., & Pierrard, R. Optimum-Benchmark: A framework for benchmarking performance of Transformers models[CP/OL]. GitHub, 2023. Retrieved from https://github.com/huggingface/optimum-benchmark.

[10] Beeching, E., Fourrier, C., et al. Open LLM Leaderboard v2[EB/OL]. Hugging Face, 2024. Retrieved from https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

[11] Guo, C., Pleiss, G., Sun, Y., et al. On calibration of modern neural networks[C]. In: Proceedings of the 34th International Conference on Machine Learning (ICML 2017), 2017: 1321–1330. DOI: 10.48550/arXiv.1706.04599.

[12] Naeini, M.P., Cooper, G., & Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning[C]. In: Proceedings of the 29th AAAI

Conference on Artificial Intelligence (AAAI 2015), 2015: 2901–2907. DOI: 10.5555/2886521.2886722.

[13] Williams, S., Waterman, A., & Patterson, D. Roofline: An insightful visual performance model for multicore architectures[J]. Communications of the ACM, 2009, 52(4): 65–76. DOI: 10.1145/1498765.1498785.

[14] Kaplan, J., McCandlish, S., Henighan, T., et al. Scaling laws for neural language models[EB/OL]. arXiv:2001.08361, 2020. Retrieved from https://arxiv.org/abs/2001.08361.

[15] Hoffmann, J., Borgeaud, S., Mensch, A., et al. Training compute-optimal large language models[EB/OL]. arXiv:2203.15556, 2022. Retrieved from https://arxiv.org/abs/2203.15556.

[16] Shao, J. Bootstrap sample size in nonregular cases[J]. Proceedings of the American Mathematical Society, 1994, 122(4): 1251–1262. DOI: 10.1090/S0002-9939-1994-1219710-7.

Note: Figure translations are in progress. See original paper for figures.