Quantifying Ranking Uncertainty in LLM Benchmarks

Bitya Neuhof [1] **Yuval Benjamini** [1]

Abstract

Pretrained models are typically ranked on multi-task leaderboards to assess their effectiveness across diverse tasks. Rank confidence intervals were recently introduced as a method to quantify the uncertainty in these rankings by aggregating pairwise hypothesis tests. In this work, we analyze the sources of uncertainty in the knowledge evaluation benchmark MMLU and show how hypothesis tests can be modified to account for their effects. We demonstrate that ranking variability across MMLU subjects is substantial and should be considered when comparing LLMs or identifying the top-performing models.

1. Introduction

The practice of model ranking has been formalized through public leaderboards, such as the Hugging Face Open LLM Leaderboard (Fourrier et al., 2024), which provides an overview of model rankings across diverse tasks. As a result, leaderboards have become the primary way researchers report empirical progress, with rankings signaling advancement and fostering competition. For each benchmark, models are evaluated on multiple tasks and prompts. Because evaluation metrics use different numerical scales (Demšar, 2006; Longjohn et al., 2025), a summary performance metric is typically ordered and presented as a ranking. These rankings then guide model comparison and selection. Despite the widespread adoption of rankings, important limitations remain. Leaderboard results can vary significantly depending on the evaluation set and the consistency of evaluation methods (Rising, 2021). Moreover, rankings are typically presented as single values, without any quantification of their uncertainty (Miller, 2024; Ackerman et al., 2025). This lack of uncertainty quantification can obscure substantial variation in model performance across tasks or prompts, especially for large, general-purpose pretrained models (such as LLMs). While confidence intervals (CIs) for model scores are common, CIs for their ranks are needed to support the identification of significant model differences (Foucart et al., 2025; Valdeira & Soares, 2025).

In this paper, we discuss how directional pairwise tests and rank CIs can improve uncertainty estimation for ranking. We demonstrate the role of tests on the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021; Polo et al., 2024), which consists of multiple-choice questions across a wide range of subjects. Our focus is on the choice of hypothesis tests underlying the construction of the rank CIs on the resulting intervals. Using these tools, we identify sources of ranking uncertainty, show that variability between subjects exceeds that between prompt variants, and highlight the importance of subject-level analysis for detecting performance differences between models. The code for our analyses is publicly available at quantifying-rank-uncertainty.

2. Rank Confidence Intervals

In this section, we describe a method for constructing rank confidence intervals (CIs) from paired samples (Holm, 2013; Al Mohamad et al., 2021; Neuhof & Benjamini, 2024; Foucart et al., 2025). Rank CIs provide a principled way to communicate the uncertainty in model rankings, which is not captured by reporting only raw ranks. Here, we treat the sources of uncertainty or levels of aggregation as given. In the next section, we examine the influence this choice has on how the results are interpreted.

2.1. Terminology and Definitions

Let $\mathcal{M}$ be a set of models, with $N_{\mathcal{M}}$ models. For simplicity, assume a balanced design, meaning that for each model m_j we observe a vector of per-unit performance scores s_{ij} for units y_i, $i = 1, \ldots, n$. Higher scores mean better performance. The observed model scores $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_{N_{\mathcal{M}}})$ are obtained by averaging over the units $\hat{\mu}_j = \frac{1}{n}\sum s_{ij}$. Raw observed ranks $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_{N_{\mathcal{M}}})$, with $\hat{r}_j \in \{1, \ldots, N_{\mathcal{M}}\}$, are assigned by ordering $\hat{\mu}$ so the highest score gets rank $N_{\mathcal{M}}$ and the lowest gets rank 1.

Let μ_j denote the true performance score of model m_j, and $\hat{\mu}_j$ denote its noisy estimate. The true rank of model m_j

[1]Department of Statistics and Data Science, The Hebrew University of Jerusalem, Jerusalem, Israel. Correspondence to: Bitya Neuhof <bitya.neuhof@mail.huji.ac.il>, Yuval Benjamini <yuval.benjamini@mail.huji.ac.il>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

is determined by the relative position of μ_j among the true scores $(\mu_1, \ldots, \mu_{N_\mathcal{M}})$. When there are no ties, the true rank of m_j can be written in two equivalent ways:

$$\begin{aligned} r_j &= 1 + \#\{\mu_k > \mu_j : k \neq j\} \\ &= N_\mathcal{M} - \#\{\mu_k < \mu_j : k \neq j\} \end{aligned} \tag{1}$$

The first expression counts how many models outperform m_j; the second counts how many are outperformed by m_j. This equivalence enables the construction of rank CIs using directional pairwise comparisons: instead of the exact true scores, we rely on statistical evidence to identify models that are significantly better or worse than m_j. These identified groups provide upper and lower bounds on the possible rank of m_j. Rank intervals can be used to communicate the uncertainty in the true ranks. Let $[L_j, U_j]$ be the rank CI for model m_j, with α as the miscoverage rate. Here, r_j denotes the true rank of model m_j. If there are ties in the true scores, r_j may be a set of possible ranks rather than a single value. In this case, the coverage definition below treats r_j as a set and checks whether it is contained in the interval.

Definition 2.1. (Valid Rank CI) A rank interval for model m_j is valid if the probability that the true rank r_j is not contained in the interval $[L_j, U_j]$ is at most α.

$$\mathbb{P}\big(r_j \subseteq [L_j, U_j]\big) \geq 1 - \alpha. \tag{2}$$

A set of valid rank CIs may guarantee coverage for each model individually (marginal coverage) or for all models simultaneously (simultaneous coverage). See Appendix A for formal definitions.

We construct rank CIs using pairwise location tests, extending the standard approach of sorting raw scores. This method allows control over ranking errors, either at the individual model level or across all models (Holm, 2013; Al Mohamad et al., 2022). For each pair of models, with true scores μ_j and μ_k, we test two one-sided hypotheses:

$$\begin{aligned} &H_{jk;0} : \mu_j \leq \mu_k \text{ vs } H_{jk;1} : \mu_j > \mu_k, \text{ and} \\ &H_{kj;0} : \mu_k \leq \mu_j \text{ vs } H_{kj;1} : \mu_k > \mu_j. \end{aligned} \tag{3}$$

The result of each hypothesis test is a p-value p_{jk} for every ordered pair of models (m_j, m_k).

For $N_\mathcal{M}$ models, there are $N_\mathcal{M}(N_\mathcal{M} - 1)$ such pairwise comparisons. Rank CIs for the true ranks are obtained by *counting* rejections of hypotheses after controlling for family-wise error rate (FWER). FWER control can be performed for decisions associated with a single model at a time (Holm, 2013; Mogstad et al., 2024) or simultaneously across all models (Al Mohamad et al., 2022; Neuhof & Benjamini, 2024).

Given a set of hypotheses $\{H_{jk;0}, j \neq k\}$, the lower and upper bounds of the ranks are:

$$\begin{aligned} L_j &= 1 + \#\{k \neq j : H_{jk;0} \text{ is rejected}\}, \\ U_j &= M - \#\{k \neq j : H_{kj;0} \text{ is rejected}\}. \end{aligned} \tag{4}$$

Thus, the lower and upper rank bounds are obtained by applying the same logic in opposite directions: significant evidence that other models are better moves U_j downward to a worse (lower) rank, while significant evidence that other models are worse moves L_j to a better (higher) rank. In our implementation, we control for multiple comparisons using Holm's correction (Holm, 1979). The algorithms to construct rank CIs from the pairwise comparisons, as well as proofs of validity, are detailed in Appendix A.

2.2. Statistical Choices in Balanced Designs

The construction of rank CIs from pairwise location tests leaves central statistical questions: which hypotheses to test, which population to consider, and which error to control. These choices correspond to distinct probability spaces and yield distinct interpretations of the resulting rank intervals. Importantly, these abstract decisions directly determine the statistical procedures and the interpretation of the results; for example, the choice of population affects which units are compared, while the error-control target shapes the confidence guarantees. The rest of the paper uses this distinction to clarify which kind of ranking uncertainty is being quantified and why the choice of hypothesis-testing procedure and error-control target is not merely a technical detail but an integral part of the inferential claim being made. We demonstrate the different forms of uncertainty and how they translate to pairwise comparisons through the paired t-test. In a balanced design with approximately normally distributed data, define the p-value p_{jk} as $p_{jk} = T_{n-1}(t)$, where $t = \frac{\bar{d}}{\sqrt{n} sd(\bar{d})}$. Here, $\bar{d}$ is the difference of the sample means, n is the sample size, and $sd(\bar{d})$ is the sample standard deviation (SD) of the difference. T_{n-1} denotes the cumulative distribution function (CDF) of the T-distribution with $n - 1$ degrees of freedom. Robust alternatives use a variant of t and estimate its distribution from the sample, for example, using bootstrap or permutation methods (Wilcox, 2023); more details and an example are in Appendix A and 5. We use n for the number of units being compared, but n varies across definitions of units. In unbalanced designs, a similar approach can also be applied with a random-effects or mixed-effects model (Bates et al., 2015).

3. Rank CIs Application to MMLU

The MMLU benchmark (Hendrycks et al., 2021) evaluates LLMs' knowledge and problem-solving abilities across a wide range of subjects, spanning STEM, humanities, and social sciences, via multiple-choice questions. We use the PromptEval extension (Polo et al., 2024) of MMLU, which

provides a correctness matrix covering 100 prompt variations, 15 models, and 57 subjects, each with at least 100 questions. The data is publicly available (MIT license) to download from Hugging Face [1]. PromptEval was designed to evaluate model sensitivity to prompt variation, introducing two main sources of variability for each subject: across prompts and across questions. In addition, we expect variability between subjects and model correlations that may differ by subject. In a complex benchmark like MMLU, relying on a single average accuracy across subjects risks obscuring both between-subject and within-subject variability. Such variability arises due to differences in subject domains, question content, and the number of questions per subject. To make subjects comparable, we randomly sample 100 questions from each subject.

We measure variability by constructing rank CIs for models, using the hypothesis tests described in Section 2 (Equation 3). Let $\mathcal{M}$, $\mathcal{T}$, $\mathcal{Q}(\mathcal{T})$, and $\mathcal{V}$ denote the sets of models, subjects, questions, and prompt variants, respectively. Define $S : (\mathcal{M}, \mathcal{T}, \mathcal{Q}(\mathcal{T}), \mathcal{V}) \to \{0, 1\}$ as a binary function that assigns a score of 1 for correct answers and 0 for incorrect ones. Let s_{ij} denote the score of model m_j and unit y_i, where each unit may correspond, for example, to a single question-prompt pair for a given subject, or to an aggregation over multiple questions and prompts, depending on the analysis context. This definition of a unit forms the basis for our hypothesis tests and for constructing the vector of differences between the models analyzed in this paper. The total number of units, n, constitutes the population for the paired t-test. All reported score and rank CIs use a miscoverage level of $\alpha = 0.05$. Note that all the rank CIs in this section have marginal coverage; an example of simultaneous rank CIs is provided in Appendix B.2.

3.1. Benchmark-Level Variability

A naive way to compare two models is to treat each individual observation as a separate unit. Specifically, a unit y_i is indexed by all unique triples (t_b, q_{bc}, v_l), where t_b is a subject, q_{bc} is a question for subject t_b, and v_l is a prompt variant. For a pair of models (m_j, m_k), we form:

$$d_{[jk,b,bc,l]} = S(m_j, t_b, q_{bc}, v_l) - S(m_k, t_b, q_{bc}, v_l), \quad (5)$$

and test whether the average of these unit-level differences is zero. For models (m_j, m_k), the estimator is

$$\bar{d} = \frac{1}{N_{\mathcal{T}} N_{\mathcal{Q}(\mathcal{T})} N_{\mathcal{V}}} \sum_b \sum_c \sum_l d_{[jk,b,bc,l]} = \hat{\mu}_j - \hat{\mu}_k \quad (6)$$

where $N_{\mathcal{T}}$ is the number of subjects, $N_{\mathcal{Q}(\mathcal{T})}$ is the number of questions per subject, and $N_{\mathcal{V}}$ is the number of prompt variants. The sample size is therefore the number of units $n = N_{\mathcal{T}} N_{\mathcal{Q}(\mathcal{T})} N_{\mathcal{V}}$, and the SD measures the spread of $d_{[jk,b,bc,l]}$ around $\bar{d}$.

In Figure 1 we show the CIs for the scores, alongside the rank CIs obtained from the differences described by Equation 5.

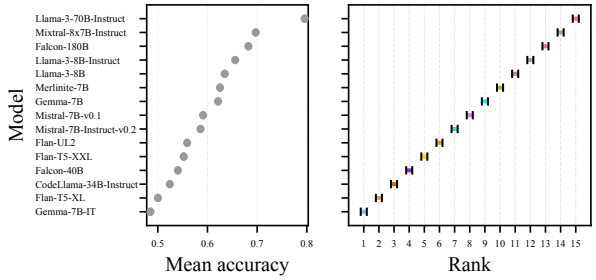


Figure 1. Scores with 95% CIs and rank CIs across all subjects, prompt variants, and questions. The models are clearly distinguishable, as indicated by non-overlapping CIs.

However, this test only supports inference for the specific evaluation set, not for new or unseen data. Treating all $d_{[jk,b,bc,l]}$ as independent observations would generally underestimate uncertainty, because it ignores dependencies within subjects, questions, or prompts; this approach reflects only the sampling variation over the observed unit-level differences. It does not account for variation that would arise from sampling new subjects, questions, or prompt variants; thus, inference is restricted to the specific evaluation set at hand. As a result, the uncertainty measured here provides a narrow view, which rarely aligns with the broader interpretation desired when analyzing benchmark results.

Subject Heterogeneity A better approach is to treat the subjects as a sample from a subject population. That is, each unit consists of a subject t_b, and an average score (accuracy) over all questions and all prompt variants of that subject:

$$d_{[jk,b,\cdot,\cdot]} = \frac{1}{N_{\mathcal{Q}(\mathcal{T})} N_{\mathcal{V}}} \sum_c \sum_l d_{[jk,b,bc,l]}. \quad (7)$$

For models (m_j, m_k), we have the same t-statistic numerator $\bar{d}$ as in Equation 6, but here the SD quantify the spread of $d_{[jk,b,\cdot,\cdot]}$ around $\bar{d}$, and the number of units is the number of subjects $n = N_{\mathcal{T}}$. The test supports extrapolation to new subjects and their associated questions, whereas the prompt variants are considered fixed and averaged are averaged within each subject. In Figure 2, we see that model rankings differ between subjects. The hypotheses are tested with respect to $(\mu_j^{\text{subject}}, \mu_k^{\text{subject}})$, where:

$$\mu_j^{\text{subject}} = \mathbb{E}_{T,Q(T)} \left[\frac{1}{N_V} \sum_l [\mathbb{E}_S[S(m_j, T, Q(T), v_l)] \right]. \quad (8)$$

[1] https://huggingface.co/datasets/PromptEval/PromptEval_MMLU_correctness

The expectation is taken over a population of subjects ($\mathcal{T}$), questions sampled within subjects ($\mathcal{Q}(\mathcal{T})$), and the randomness of the score function (S).

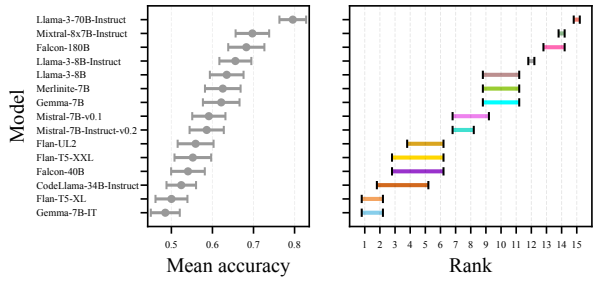


Figure 2. Scores with 95% CIs and rank CIs from accuracy scores per subject. The overlaps between models' intervals indicate, for example, that 3 models can be ranked as the fourth from the top.

Sensitivity to Prompt Variants A second analysis compares the average effect of each prompt across all subjects. In this analysis, each prompt serves as the inferential unit. We define

$$d_{[jk,\cdot,\cdot,l]} = \frac{1}{N_{\mathcal{T}} N_{\mathcal{Q}(\mathcal{T})}} \sum_b \sum_c d_{[jk,b,bc,l]}. \tag{9}$$

The t-test is based on the variance of $\bar{d}$ (see Equation 6), calculated across prompts, where the number of units n equals the number of prompt variants ($N_{\mathcal{V}}$). The variance is computed across these prompt-level units and quantifies how model differences between models vary with different prompt formulations. The estimand is:

$$\mu_j^{\text{prompt}} = \mathbb{E}_V \left[\frac{1}{N_{\mathcal{T}}} \sum_b \frac{1}{N_{\mathcal{Q}(\mathcal{T})}} \sum_c \mathbb{E}_S[S(m_j, t_b, q_{bc}, V)] \right]. \tag{10}$$

Figure 3 reveals a pattern similar to Figure 1, suggesting that the number of prompts is large enough, relative to their variability, to yield stable model rankings for the observed subjects. The results are based on $N_{\mathcal{V}} = 100$ prompt variants. To determine whether the number of units ($N_{\mathcal{T}} = 57 < N_{\mathcal{V}}$) affects the results compared to those in Figure 2, we randomly sampled $N_{\mathcal{T}}$ prompt variants from the total $N_{\mathcal{V}}$ and recalculated the rank CIs. Consistent results after subsampling suggest that subject-level variability dominates over prompt-level variability.

For the three analyses described above, the numerator of the t-statistic is identical ($\bar{d}$); the deviation in the denominator ($sd(\bar{d})$), as well as the degrees of freedom ($(n-1)$), are different.

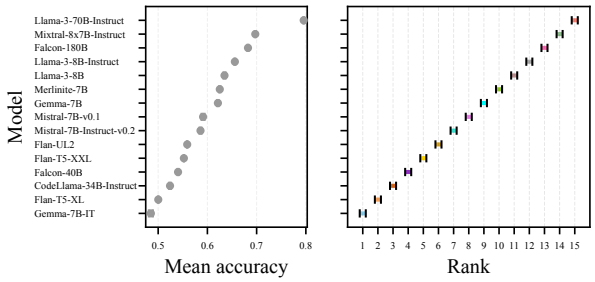


Figure 3. Scores with 95% CIs and rank CIs from accuracy scores per prompt variant.

3.2. Subject Level Variability

Here, we construct rank CIs for each subject independently, without regard to other subjects. We show examples of two subjects, *clinical knowledge* and *virology*. We change the ranking population to a specific subject at a time. The t-test numerator for models (m_j, m_k) is subject-specific, representing the average difference in model accuracy for a given subject:

$$\bar{d}^b = \hat{\mu}_j^b - \hat{\mu}_k^b = \frac{1}{N_{\mathcal{Q}(t_b)} N_{\mathcal{V}}} \sum_c \sum_l d_{[jk,b,bc,l]}. \tag{11}$$

Within each subject t_b, we can further compare models based on their average performance, either across all questions (aggregating over prompts) or across all prompts (aggregating over questions).

For the question-based comparison within a subject, we define:

$$d_{[jk,b,bc,\cdot]} = \frac{1}{N_{\mathcal{V}}} \sum_l d_{[jk,b,bc,l]}. \tag{12}$$

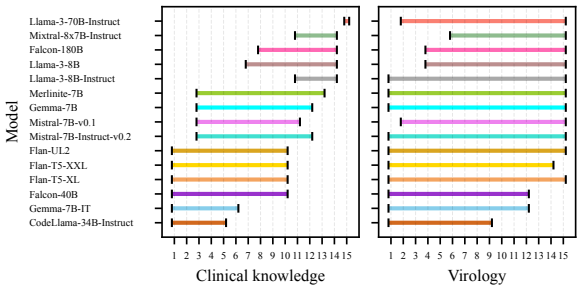
Here, paired differences are aggregated at the question level, and variance is assessed by how much $d_{[jk,b,bc,\cdot]}$ deviates from $\bar{d}^b$.

For the prompt-based comparison within a subject, we define:

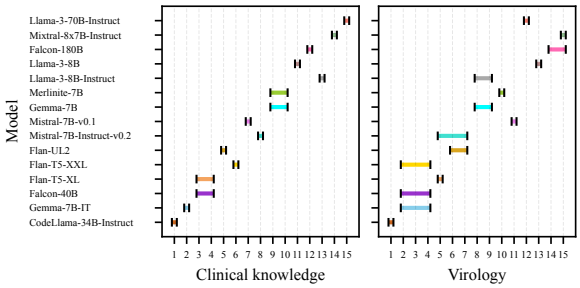
$$d_{[jk,b,\cdot,l]} = \frac{1}{N_{\mathcal{Q}(t_b)}} \sum_c d_{[jk,b,bc,l]}. \tag{13}$$

In this case, paired differences are aggregated at the prompt level, and the t-test variance is based on the variance of $\bar{d}^b$ calculated across prompts.

Figure 4 presents per-subject rank CIs, calculated according to the differences defined in Equations 12 and 13. Although many intervals overlap for both subjects, in clinical knowledge, several distinct groups of models can be identified. Both the number of questions and the number of prompt variants are 100 ($N_{\mathcal{Q}(t_b)} = N_{\mathcal{V}}$). As a result, differences between the rank CIs of two models directly reflect whether



(a) Rank CIs using questions as units.



(b) Rank CIs using prompt variations as units.

Figure 4. Subject-specific rank CIs with aggregation across questions (a) and prompts (b); the variability between questions exceed that between prompts. A model can be ranked clearly as the best for one subject (left), but not for another (right).

one model is statistically superior to, or equivalent with, another.

These two analyses use the same observed scores but address different inferential questions. The first analysis extrapolates model performance over questions within a fixed subject, while the second extrapolates over prompts within that subject.

4. Hypothesis Test Specification

In this section, we elaborate on additional specifications of the pairwise comparison process that influence the resulting rankings. We introduce the main concepts and approaches here; examples of their implementation on the MMLU benchmark can be found in Appendix B.

Statistical Difference vs Meaningful Difference Figure 1 presents the mean accuracy and corresponding rank CIs for all models. The spread of accuracy among middle-ranked models is quite small; while the models differ significantly in the pairwise hypothesis tests, yielding rank CIs that contain only a single rank. The actual differences in performance may not be meaningful in practice. To address this, we can incorporate an indifference margin into the hypothesis specification (Gibbons et al., 1979). This approach tests whether the difference between models exceeds a pre-specified margin, rather than testing whether there is any difference. Formally, let $\delta > 0$ be an indifference margin, We test the following hypotheses:

$$\begin{aligned} &H^{\delta}_{jk;0} : \mu_j - \mu_k \leq \delta \text{ vs } H_{jk;1} : \mu_j - \mu_k > \delta, \text{ and} \\ &H^{\delta}_{kj;0} : \mu_k - \mu_j \leq \delta \text{ vs } H_{kj;1} : \mu_k - \mu_j > \delta. \end{aligned} \tag{14}$$

When $\delta = 0$, these hypotheses are identical to those defined by Equation 3. For example, in the context of the MMLU benchmark, setting $\delta = 0.02$ means that if the accuracy difference between two models is less than 2%, they are considered practically equivalent for real-world applications; that is, such a difference is not considered meaningful in practice. Results for this example are in Appendix B.3.

The indifference margin can be set in advance using domain knowledge or determined empirically from the distribution of performance scores. If chosen empirically, the margin should be selected using sample splitting or alpha splitting (Hansen, 2000; Huang et al., 2022).

Non Informative Units In Section 3, we presented both the benchmark-level rank CIs and the per-subject rank CIs. Per-subject rank CIs can be used to improve the interpretability of benchmark rank CIs by indicating which subjects are not informative, in other words, those that do not contribute information about the differences between models. For example, in the left panel of Figure 4(a), we can see that for virology, all models are almost equivalent, meaning there is no significant difference in their performance for this subject. To select a subset of subjects, we can split the data or the α, construct per-subject rank CIs for all models, filter out non-informative subjects, and then construct the benchmark-level rank CIs based only on the informative ones. Moreover, identifying subjects with substantial differences between models is valuable, as this information can help select a model for a new subject. See an example in Appendix B.4.

5. Conclusions

We emphasize the advantages of rank CIs over CIs for model scores, particularly for quantifying uncertainty in aggregated model performance. By exploring different ways to represent and aggregate model performance, we demonstrate how ranking based on hypothesis tests provides a clear framework for defining and measuring ranking variability and correlations between models. These definitions clarify the sources of uncertainty that are captured by the rank CIs.

In this paper, we analyzed ranking uncertainty using a single benchmark. However, LLM leaderboards typically include multiple, diverse benchmarks. In future work, we will address how to combine insights from individual benchmarks to summarize uncertainty across multiple benchmarks.

Acknowledgments

This work was supported by the Israel Science Foundation.

Impact Statement

Leaderboards are used by the Machine Learning community to showcase model performance. This paper shows that rankings of models within the benchmark can vary due to multiple sources of uncertainty, and that this variation should inform a more transparent and nuanced view of machine learning model performance.

References

Ackerman, S., Farchi, E., Raz, O., and Toledo, A. Statistical multi-metric evaluation and visualization of llm system predictive performance. *arXiv preprint arXiv:2501.18243*, 2025.

Al Mohamad, D., van Zwet, E., Solari, A., and Goeman, J. Simultaneous confidence intervals for ranks using the partitioning principle. *Electronic Journal of Statistics*, 15 (1):3109–3134, 2021. doi: 10.1214/21-EJS1847.

Al Mohamad, D., Goeman, J. J., and van Zwet, E. W. Simultaneous confidence intervals for ranks with application to ranking institutions. *Biometrics*, 78(1):238–247, 2022.

Bates, D., Mächler, M., Bolker, B., and Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01. URL https://www.jstatsoft.org/index.php/jss/article/view/v067i01.

Benjamini, Y. and Yekutieli, D. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

Chandra, O. and Xie, M.-g. Finite-sample valid rank confidence sets for a broad class of statistical and machine learning models. *arXiv preprint arXiv:2512.00316*, 2025.

Chetverikov, D., Mogstad, M., Morgen, P., Romano, J., Shaikh, A., and Wilhelm, D. csranks: an r package for estimation and inference involving ranks. *arXiv preprint arXiv:2401.15205*, 2024.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7 (Jan):1–30, 2006.

Foucart, A., Elskens, A., and Decaestecker, C. Ranking the scores of algorithms with confidence. In *ESANN 2025 proceedings*, pp. 431–436, 2025.

Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., and Wolf, T. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

Gibbons, J. D., Olkin, I., and Sobel, M. An introduction to ranking and selection. *The American Statistician*, 33(4): 185–195, 1979.

Hansen, B. E. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603, 2000.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Holm, S. Confidence intervals for ranks. Technical report, Uppsala University, 2013.

Huang, C., Li, P., and Martin, C. R. Simplification or simulation: Power calculation in clinical trials. *Contemporary Clinical Trials*, 113:106663, 2022.

Longjohn, R., Gopalan, G., and Casleton, E. Statistical uncertainty quantification for aggregate performance metrics in machine learning benchmarks. *arXiv preprint arXiv:2501.04234*, 2025.

Miller, E. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Mogstad, M., Romano, J. P., Shaikh, A. M., and Wilhelm, D. Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *The Review of Economic Studies*, 91(1):476–518, 01 2024. ISSN 0034-6527. doi: 10.1093/restud/rdad006.

Neuhof, B. and Benjamini, Y. Confident feature ranking. In *International Conference on Artificial Intelligence and Statistics*, pp. 1468–1476. PMLR, 2024.

Polo, F. M., Xu, R., Weber, L., Silva, M., Bhardwaj, O., Choshen, L., de Oliveira, A. F., Sun, Y., and Yurochkin, M. Efficient multi-prompt evaluation of llms. *Advances in Neural Information Processing Systems*, 37:22483–22512, 2024.

Rising, J. Uncertainty in ranking. *arXiv preprint arXiv:2107.03459*, 2021.

Valdeira, F. and Soares, C. Ranking with confidence for large scale comparison data. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*, pp. 223–232. SIAM, 2025.

Wilcox, R. R. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

Wilcox, R. R. Comparing two dependent groups. In *A Guide to Robust Statistical Methods*, pp. 83–96. Springer, 2023.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

A. Rank CIs definitions and methods

A.1. Marginal and Simultaneous Coverage

Our goal is to estimate the true ranks for all $N_{\mathcal{M}}$ models. Let X be an $n \times N_{\mathcal{M}}$ observed aggregation of scores so that $X \sim \mathbb{P}_\mu$. Here $\mathbb{P}_\mu$ is a distribution family indexed by the vector of means $\mu \in R^N_{\mathcal{M}}$. Let $([L_1(X), U_1(X)], \ldots, [L_{N_{\mathcal{M}}}(X), U_{N_{\mathcal{M}}}(X)])$ denote the random data-dependent rank intervals for the $N_{\mathcal{M}}$ models.

Definition A.1. (Marginal Coverage) A set of intervals has *marginal coverage* at level $1-\alpha$ if the probability that each interval cover its true rank is at least $1-\alpha$, for any μ:

$$\mathbb{P}_\mu\big(r_j \subseteq [L_j, U_j]\big) \geq 1-\alpha \quad \forall j \in \{1, \ldots, N_{\mathcal{M}}\}, \quad \forall \mu \in R^{N_{\mathcal{M}}}. \tag{15}$$

Marginal coverage does not guarantee valid selection after ranking. In many cases, simultaneous coverage is needed. CIs with simultaneous coverage remain valid under any selection (e.g., choosing a subset of models) (Benjamini & Yekutieli, 2005), but these intervals are typically wider.

Definition A.2. (Simultaneous Coverage) A set of intervals has *simultaneous coverage* at level $1-\alpha$ if the probability that all intervals cover their true ranks is at least $1-\alpha$, for any μ:

$$\mathbb{P}_\mu\big(r_j \subseteq [L_j, U_j], \quad \forall j \in \{1, \ldots, N_{\mathcal{M}}\}\big) \geq 1-\alpha, \quad \forall \mu \in R^{N_{\mathcal{M}}}. \tag{16}$$

We use Holm's procedure to adjust p-values for multiple comparisons and control the FWER. The hypotheses are defined by the desired criterion: marginal or simultaneous rank CIs. Holm's procedure controls FWER at level α, works efficiently, and is valid for any p-value dependence. Alternatively, other FWER-controlling methods, such as resampling algorithms that directly compute thresholds by resampling the data after removing mean differences (Chetverikov et al., 2024), can be used. These can yield smaller, valid intervals when sample sizes are large enough.

A.2. Constructing Rank CIs

For marginal coverage, controlling the family-wise error rate (FWER) for each model individually is sufficient (Holm, 2013). Algorithm 1 summarizes how to construct marginal rank CIs from pairwise comparisons.

Algorithm 1 Marginal Rank CI

Input:
Pairwise p-values $\{p^b_{jk} : j \neq k\}$;
FWER-controlling procedure ϕ (e.g., Holm), and miscoverage rate α_{tsk};
for $j = 1, \ldots, N_{\mathcal{M}}$ **do**
  Lower rank bound:
  Apply ϕ to $\{p^b_{kj} : k \neq j\}$ at level $\alpha_{tsk}/2$, obtaining rejection vector $R_{j,L} \in \{0,1\}^{N_{\mathcal{M}}-1}$.
  $L^b_j \leftarrow 1 + \sum_{k=1}^{N_{\mathcal{M}}-1} R_{j,L}(k)$ {Count models significantly worse than c_j}
  Upper rank bound:
  Apply ϕ to $\{p^b_{jk} : k \neq j\}$ at level $\alpha_{tsk}/2$, obtaining rejection vector $R_{j,U} \in \{0,1\}^{N_{\mathcal{M}}-1}$.
  $U^b_j \leftarrow N_{\mathcal{M}} - \sum_{k=1}^{N_{\mathcal{M}}-1} R_{j,U}(k)$ {Count models significantly better than j}
end for
Return $\big\{[L^b_1, U^b_1], \ldots, [L^b_{N_{\mathcal{M}}}, U^b_{N_{\mathcal{M}}}]\big\}$

Theorem A.3. *[Holm 2013] The rank CIs $[L^b_1, U^b_1], \ldots, [L^b_{N_{\mathcal{M}}}, U^b_{N_{\mathcal{M}}}]$ constructed by Algorithm 1 have marginal coverage rate $(1-\alpha)$.*

See Theorem 1 in Holm (2013) for the formal statement and proof. In our analyses, we use paired t-tests for significance and Holm's procedure for multiplicity control.

Theorem A.4. *[Neuhof and Benjamini 2024] The rank CIs $[L_1, U_1], \ldots, [L_{N_{\mathcal{M}}}, U_{N_{\mathcal{M}}}]$ constructed by Algorithm 2 achieve simultaneous coverage rate $(1-\alpha)$.*

See Theorem 1 in Neuhof & Benjamini (2024) for details and proof.

Algorithm 2 Simultaneous Rank CI

Input:
Pairwise p-values $\{p_{jk} : j \neq k\}$;
FWER-controlling procedure ϕ (e.g., Holm), and miscoverage rate α;
for $j = 1, \ldots, N_{\mathcal{M}}$ **do**
  Apply ϕ to all p-values at level α, obtaining rejection vector $R \in \{0,1\}^{N_{\mathcal{M}}(N_{\mathcal{M}}-1)}$.
  Lower rank bound:
  $L_j \leftarrow 1 + \sum_{k=1}^{N_{\mathcal{M}}-1} R(jk)$ {Count models significantly worse than c_j}
  Upper rank bound:
  $U_j \leftarrow N_{\mathcal{M}} - \sum_{k=1}^{N_{\mathcal{M}}-1} R(kj)$ {Count models significantly better than c_j}
end for
Return $\{[L_1, U_1], \ldots, [L_{N_{\mathcal{M}}}, U_{N_{\mathcal{M}}}]\}$

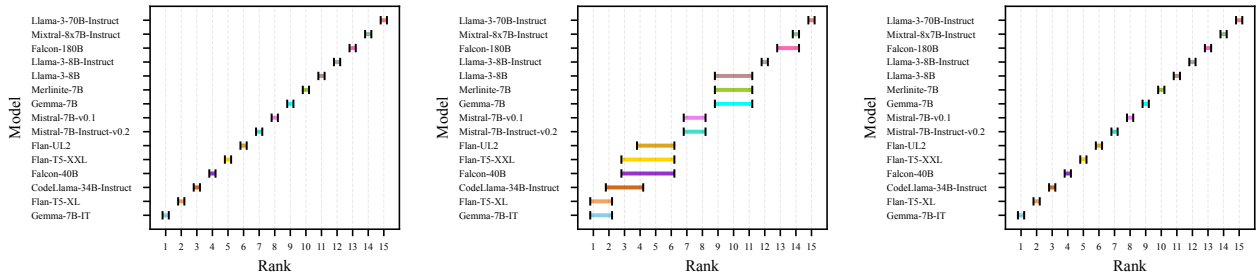
Other methods can also construct valid rank CIs with marginal or simultaneous coverage, such as those in Al Mohamad et al. (2021; 2022); Chetverikov et al. (2024); Chandra & Xie (2025); Valdeira & Soares (2025). Alternative methods for constructing rank CIs require different p-value computations or more efficient FWER procedures. For matched samples, use robust alternatives to the t-test, such as the Wilcoxon signed-rank test (Wilcoxon, 1945; Wilcox, 2011) (more robust, less power) or the trimmed-mean t-test (Wilcox, 2023) (intermediate power, requires sampling), when distributions are not normal; see an example in Appendix B.1. For non-matched samples, tests like the z-test are appropriate.

B. Additional Results

B.1. Non-parametric Paired Test

In our analysis, we use the paired t-test because we rank models based on accuracy, an average over binary scores, except for Figure 1. By the Central Limit Theorem, this average is approximately normally distributed.

However, the distribution may not be approximately normal, or even symmetric. In these cases, use a more robust test or a nonparametric test, such as the Wilcoxon signed-rank test (Wilcoxon, 1945), instead of the t-test. In Figure 5, we show benchmark-level rank CIs using the Wilcoxon test for paired comparisons instead of the t-test, as in Figures 1, 2, and 3. The rank CIs are the same as those obtained with the t-test, except for subjects as units (Figure 2).

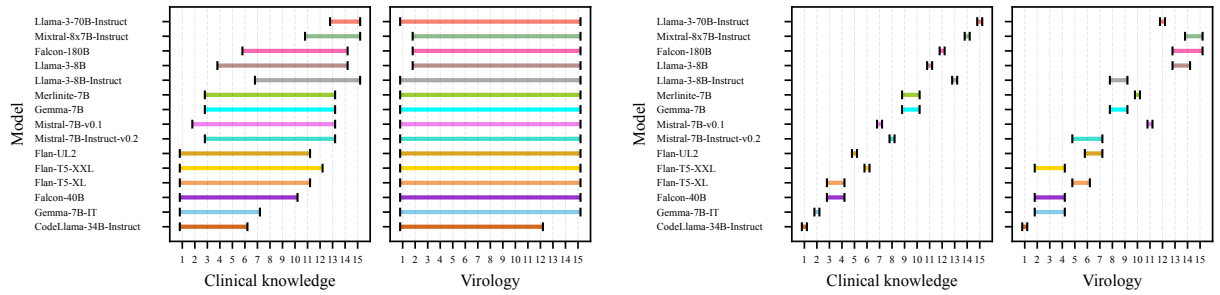


(a) Rank CIs using observations as units. *(b)* Rank CIs using subjects as units. *(c)* Rank CIs using prompt variations as units.

Figure 5. Rank CIs with Wilcoxon as the paired test and different definitions of units; observations (a), subjects (b), and prompt variants (c).

B.2. Simultaneous Rank CIs

Here, we present the results of constructing simultaneous rank CIs for the same data as in Figure 4. The simultaneous rank CIs in Figure 6 are wider than the marginal rank CIs, as they account for multiple comparisons by considering all $N_{\mathcal{M}}(N_{\mathcal{M}} - 1)$ hypotheses as a single family (see (Al Mohamad et al., 2021; Neuhof & Benjamini, 2024)). This adjustment ensures a coverage guarantee: with probability at least $1 - \alpha$, the true ranks of all selected models are covered, regardless of which subset is chosen after observing the rankings. This property is especially useful when model selection is performed based on ranking outcomes.



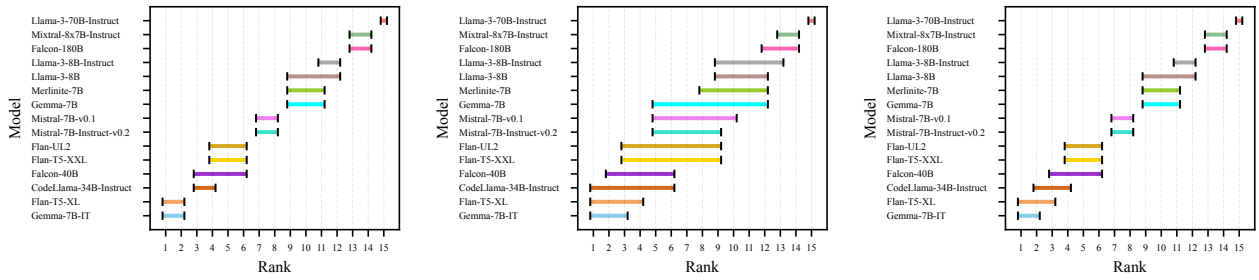
(a) Rank CIs using questions as units.

(b) Rank CIs using prompt variations as units.

Figure 6. Subject-specific rank CIs with aggregation across questions (a) and prompts (b), with simultaneous coverage.

B.3. Indifference Zone

For this demonstration, we introduce an indifference margin $\delta = 0.02$, representing a 2% difference in accuracy between models. Including this margin allows us to focus on differences that are practically meaningful rather than statistically detectable. As shown in Figure 7, the rank intervals now overlap for all definitions of units, unlike the clear separation observed when $\delta = 0$ (see Figures 1 and 3). Notably, the variability between subjects remains higher than that between prompts.



(a) Rank CIs using observations as units.

(b) Rank CIs using subjects as units.

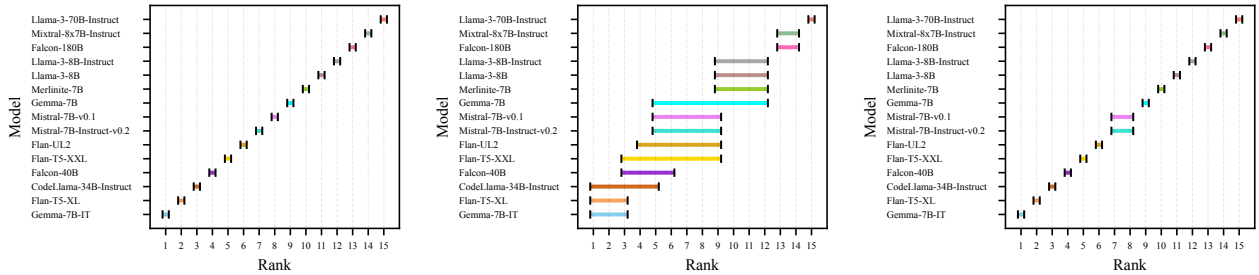
(c) Rank CIs using prompt variations as units.

Figure 7. Rank CIs with indifference zone of 2% between the accuracy of two models, and different definitions of units; observations (a), subjects (b), and prompt variants (c).

B.4. Filter Non-informative Subjects

In Figure 4, we observe that there is no difference between the models in virology. In this section, we analyze the distribution of per-subject rank CIs and identify non-informative subjects, those for which there is little or no distinguishable difference between models. Specifically, we define a non-informative subject as one in which 14 or 15 out of 15 models rank in the top 3, indicating that almost all models perform equivalently (see below for details). Using this criterion, we filtered out 27 subjects, including virology.

After filtering, we construct benchmark-level rank CIs using only the informative subjects. To maintain statistical validity, we split the miscoverage level α into two components: 0.3α for constructing per-subject rank CIs and filtering, and 0.7α for the benchmark-level rank CIs. This approach ensures appropriate control of ranking error. Figure 8 demonstrates the effect of filtering on the three sets of rank CIs: observations, subjects, and prompt variants. Filtering non-informative subjects focuses the analysis on areas where model differences are most meaningful, improving the interpretability of benchmark-level model comparisons.



(a) Rank CIs using observations as units. *(b)* Rank CIs using subjects as units. *(c)* Rank CIs using prompt variations as units.

Figure 8. Rank CIs based on a subset of 30 subjects, after filtering out subjects with almost no difference between models.