Bitya Neuhof and Yuval Benjamini

Statistics and Data Science

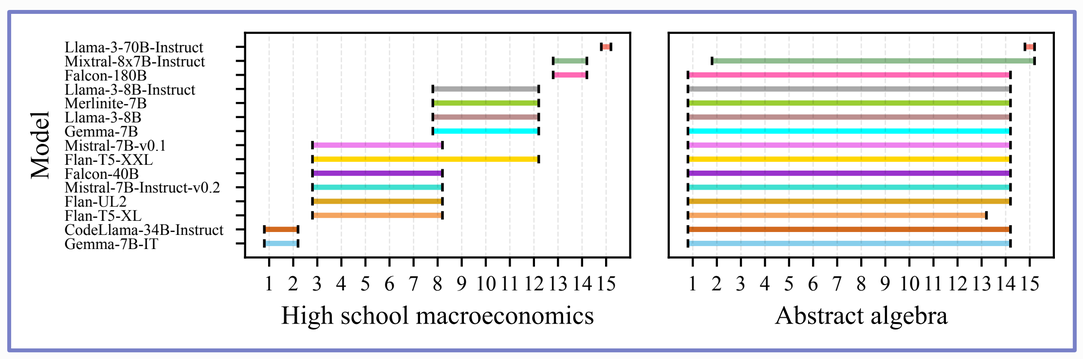
Rank Intervals for Leaderboards: A Hierarchical Framework for Model Evaluation

Summary

We construct statistically valid rank intervals at both the task and leaderboard levels to quantify uncertainty in model rankings across and within tasks.

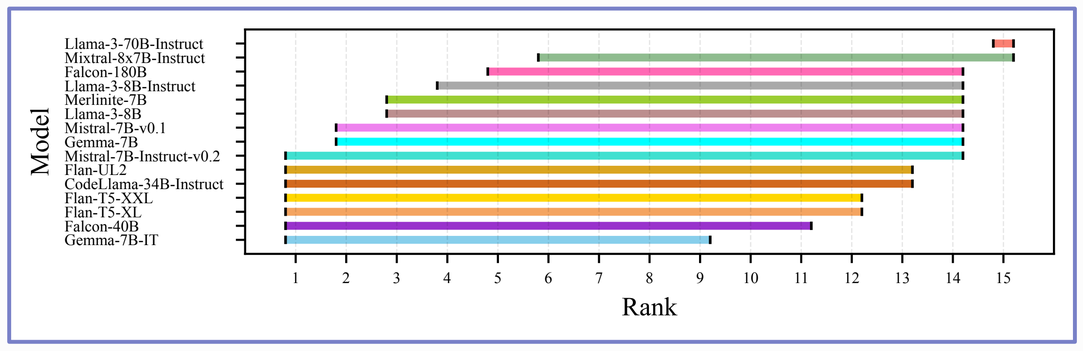
Hierarchical Ranking

Leaderboard rankings often obscure variation across tasks, and single-value summaries can hide uncertainty. The ranking by misses differences between tasks.



Rank confidence intervals (CIs) for two subjects from the MMLU benchmark (Hendrycks et al., 2021; Polo et al., 2024):

- High school macroeconomics: Narrow CIs, model groups.
- Abstract algebra: Overlap CIs, model differences unclear.



Aggregating 47 subjects (tasks) into leaderboard-level prediction intervals (PIs) summarizes model performance and quantifies expected performance on new tasks.

Task-level Rank Confidence Interval

We consider M models, $c_1, \dots, c_M$, and a task t^b.
Let μ_j^b be the true score of model c_j on task t^b.

We test all directional pairwise hypotheses:

$$(c_j, c_k)\colon\ H_{jk;0}^b\colon \mu_j^b \geq \mu_k^b \ vs\ H_{jk;1}^b\colon \mu_j^b < \mu_k^b \rightarrow \text{p-value } p_{jk}$$
$$(c_k, c_j)\colon\ H_{kj;0}^b\colon \mu_k^b \geq \mu_j^b \ vs\ H_{kj;1}^b\colon \mu_k^b < \mu_j^b \rightarrow \text{p-value } p_{kj}$$

For each model, a rank CI $[L_j^b, U_j^b]$ for the true rank r_j^b is constructed by **counting rejected hypotheses** after marginally controlling the family-wise error rate (Holm, 2013; Neuhof & Benjamini, 2024):

$$L_j^b = 1 \ + \#\{j \neq k\colon H_{jk;0}^b \ is\ rejected\}$$
$$U_j^b = M - \#\{j \neq k\colon H_{kj;0}^b \ is\ rejected\}$$

Leaderboard-level Rank Prediction Interval

Let $t^1, \dots, t^N$ be a set of exchangeable tasks, with all models evaluated on each task. Let α_{ldb} be the leaderboard-level miscoverage rate, and α_{tsk} the task-level miscoverage rate.

We construct rank-CIs with marginal coverage $1 - \alpha_{tsk}$ for each model on each task, to obtain the collection:

$$\{[L_1^b, U_1^b], \dots, [L_M^b, U_M^b]\}_{b=1}^N$$

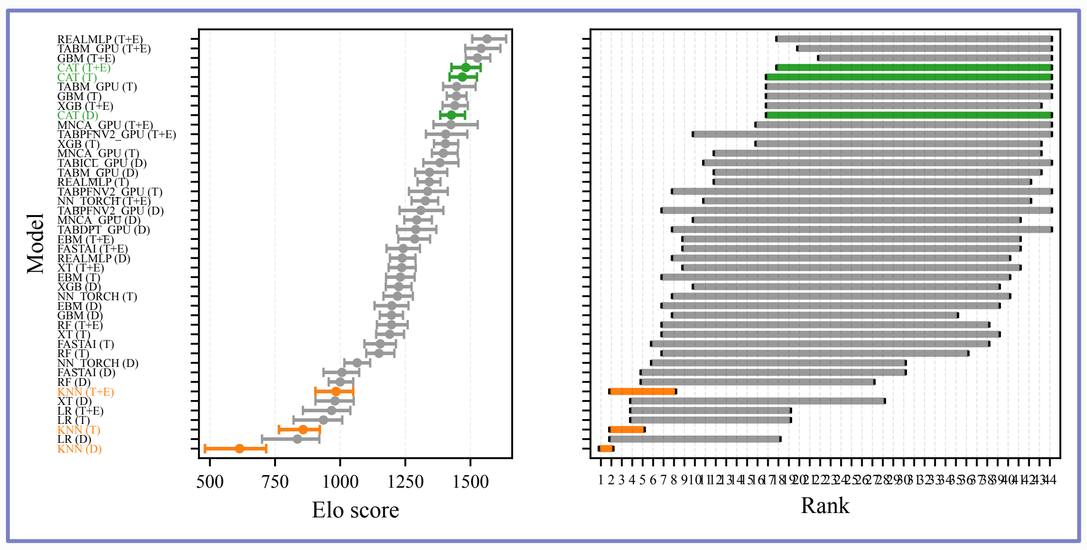
For each model c_j, we compute:

$$L_j = quantile\left(\frac{\alpha_{ldb}}{2}, \quad \{L_j^b\}_{b=1}^N\right)$$
$$U_j = quantile\left(1 - \frac{\alpha_{ldb}}{2}, \{U_j^b\}_{b=1}^N\right)$$

Theorem: Consider a new independent task t^*, The leaderboard-level rank PIs $[L_1, U_1], \dots, [L_M, U_M]$ achieve marginal task coverage of $1 - \alpha$, for $\alpha = \alpha_{ldb} + \alpha_{tsk}$.

Rank Intervals Interpretation

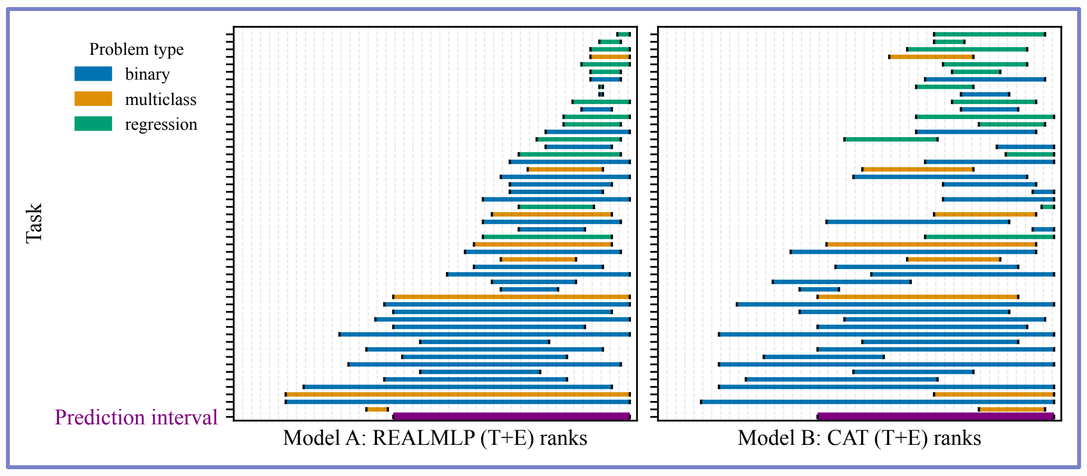
Rank intervals show when models differ or are statistically tied, and quantify ranking uncertainty.



TabArena leaderboard (Erickson et al., 2025):

- Scores with 95% CIs show average performance bounds.
- Rank PIs estimate expected performance for new tasks.

Narrow rank PIs show stable rankings. Wide PIs may result from wide task rank CIs or task heterogeneity.



Task-level rank CIs provide detailed model insights beyond leaderboard aggregation.

- CAT and REALMLP have identical rank PIs.
- The distributions of dataset-level rank CIs differ.
- CAT has narrow rank CIs for multiclass datasets.
- REALMLP's rank CIs are narrower for regression datasets.

bitya.neuhof@mail.huji.ac.il
yuval.benjamini@mail.huji.ac.il

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM
الجامعة العبرية في اورشليم القدس
Dreaming. Creating. Innovating. Since 1925