Prediction-Powered Ranking of Large Language Models

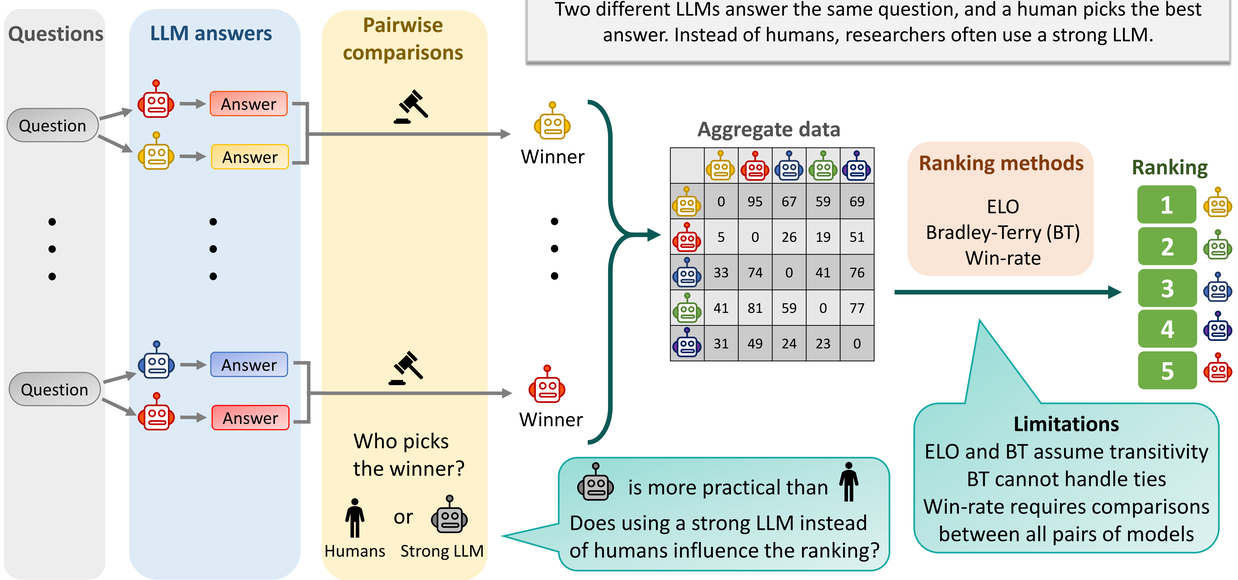
Max Planck Institute for Software Systems

Ivi Chatzi Eleni Straitouri Suhas Thejaswi Manuel Gomez-Rodriguez

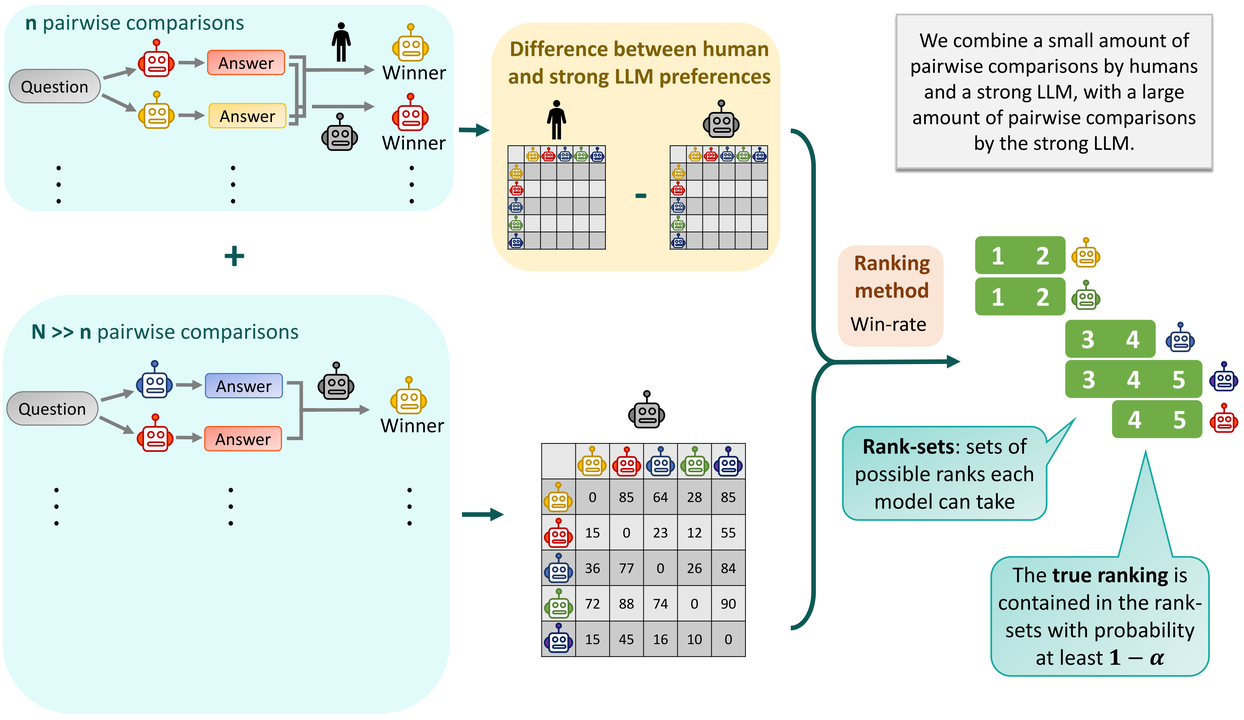
PAPER CODE

Existing methods

The goal is to rank LLMs on human alignment utilizing pairwise comparisons. Two different LLMs answer the same question, and a human picks the best answer. Instead of humans, researchers often use a strong LLM.

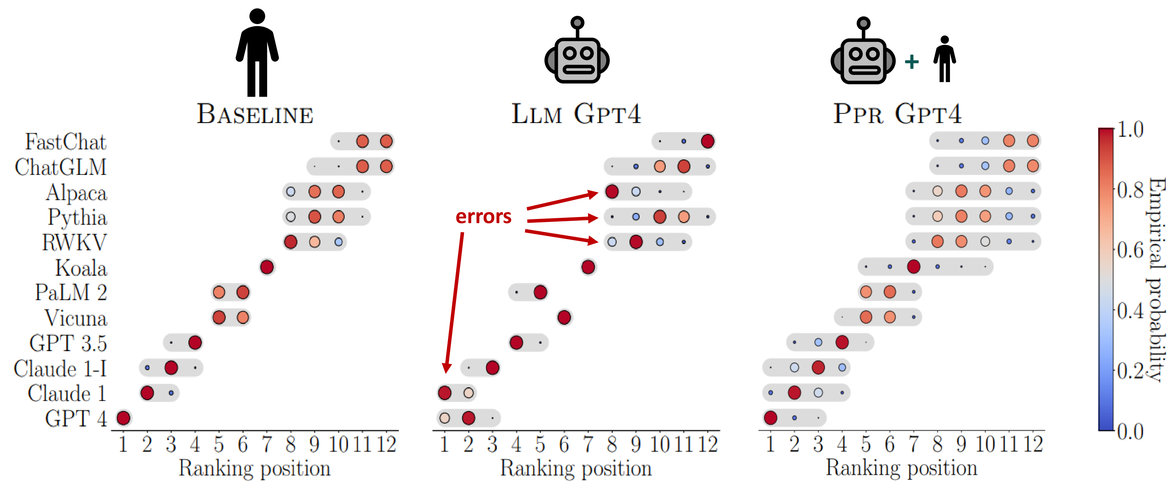


Our framework: Prediction-Powered Ranking



Rank-sets in practice

How often does the rank-set of each model contain each rank?

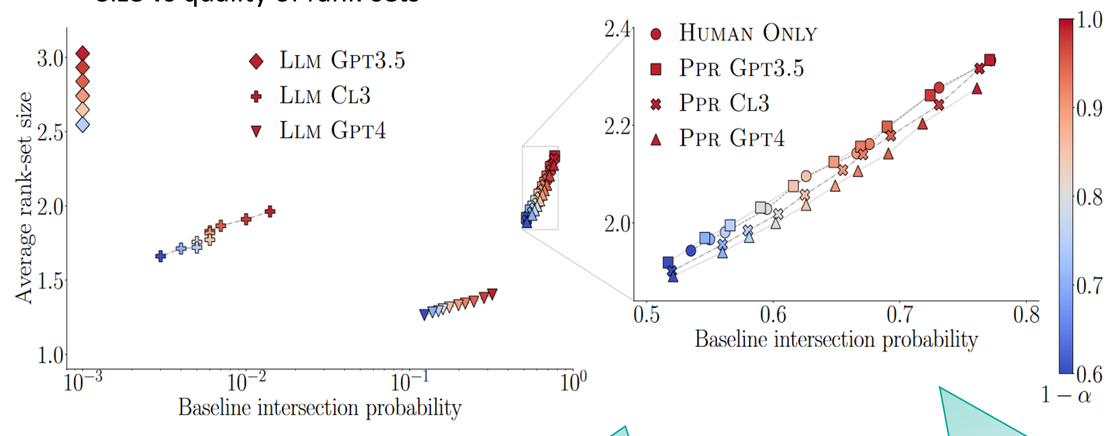


Rank-sets using all pairwise comparisons by humans

Rank-sets using only pairwise comparisons by strong LLMs have errors

Rank-sets using many pairwise comparisons by strong LLMs and a few by humans are correct but larger

Size vs quality of rank-sets



Baseline intersection probability
Empirical probability that the rank-sets of all models intersect the rank-sets using all pairwise comparisons by humans.

As the coverage increases, the rank-sets get larger

Rank-sets using PPR are smaller and more accurate than rank-sets using only a few pairwise comparisons by humans