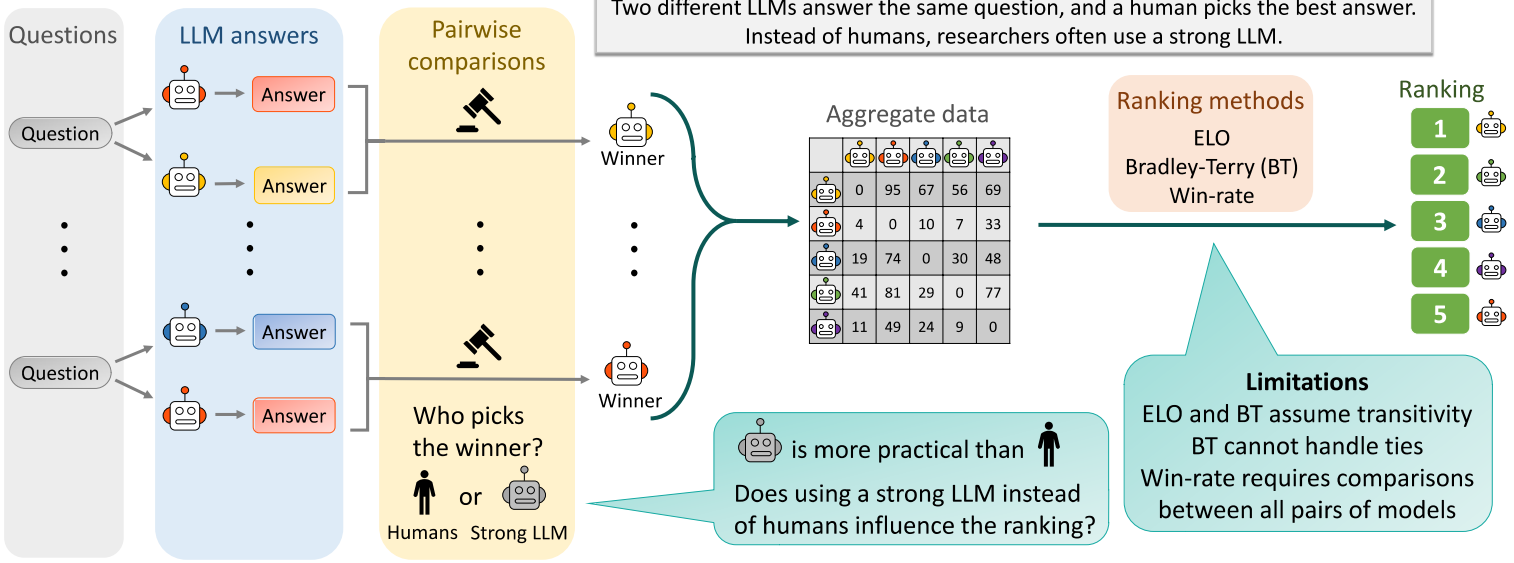
Prediction-Powered Ranking of Large Language Models

Ivi Chatzi Eleni Straitouri Suhas Thejaswi Manuel Gomez-Rodriguez

Max Planck Institute for Software Systems

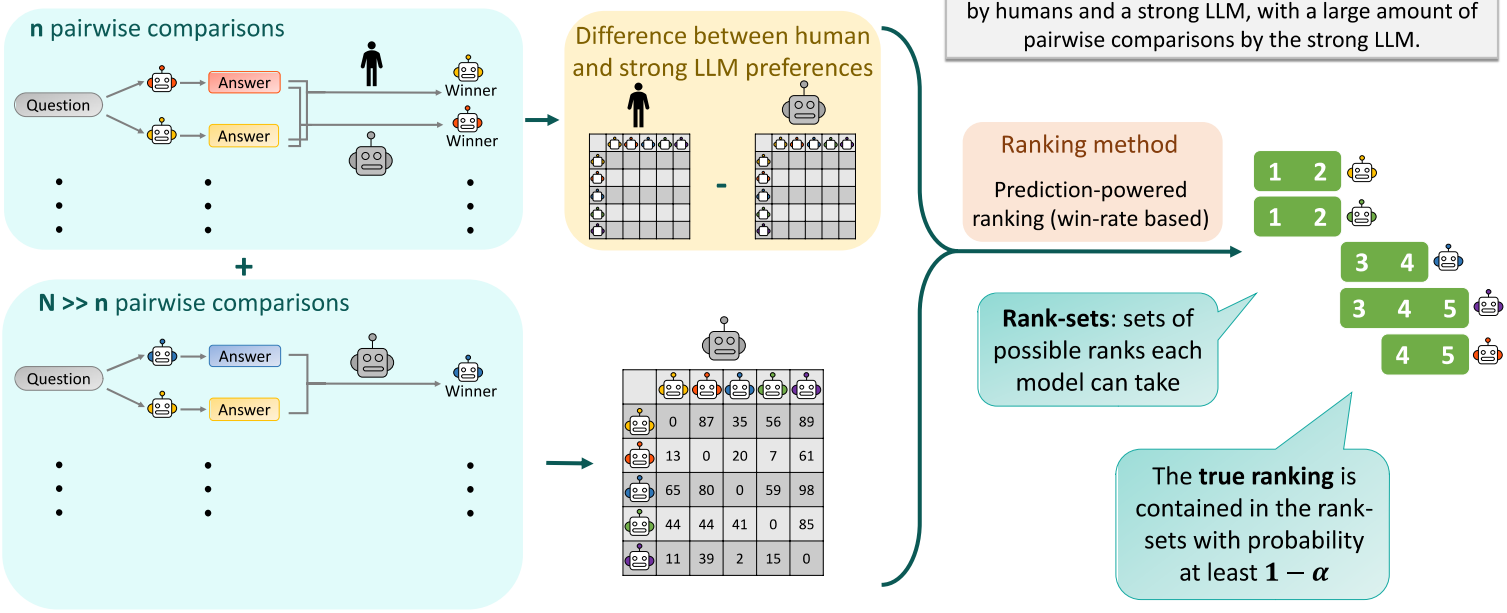
Existing methods

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.



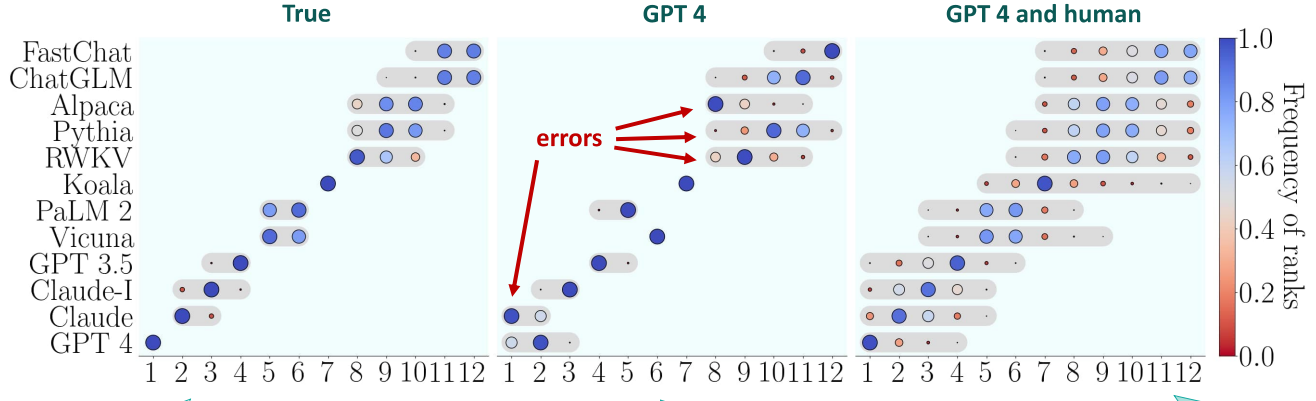
Our framework: Prediction-powered ranking

We combine a small amount of pairwise comparisons by humans and a strong LLM, with a large amount of pairwise comparisons by the strong LLM.



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 GPT 4 have errors

Rank-sets using many pairwise comparisons by GPT 4 and a few by humans are correct but larger

PAPER

CODE