Tübingen AI Center
tuebingen.ai

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

Correct Looks Better: Pairwise Comparisons Reveal Accuracy Rankings

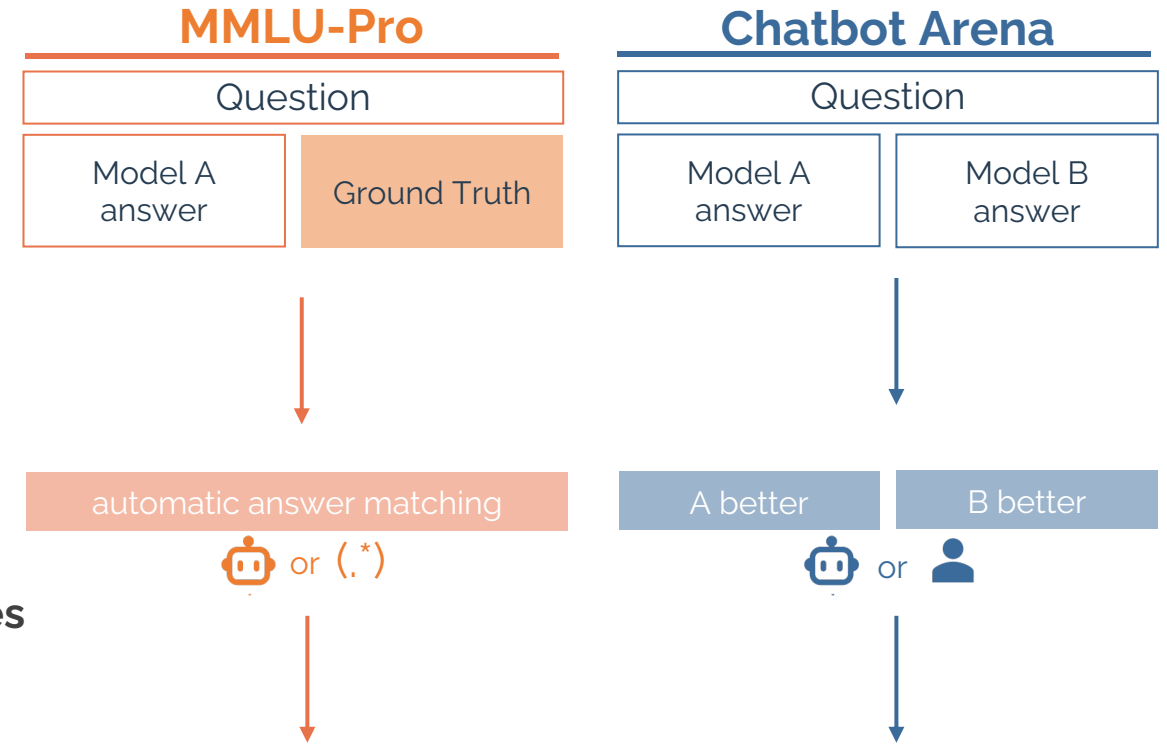
Mina Remeli, Moritz Hardt

Max Planck Institute for Intelligent Systems, Tübingen, Germany, Tübingen AI Center

Motivation

Two paradigms in model ranking

- Ranked based on:
  - **Accuracy** on a benchmark
  - Model "strength" from **pairwise preferences**



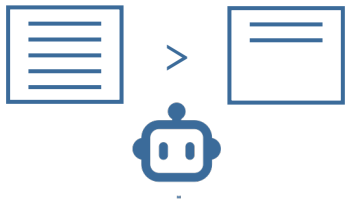
Models	Overall
Gemini-3.1-Pro	0.9116
Gemini-3-Pro(11/25)	0.901
GPT-o1	0.893
Claude-4.6-Opus(Thinking)	0.891
Gemini-3-Flash(12/25)	0.886

Model 618 / 618	Overall
claude-opus-4-6	1
claude-opus-4-6-thi…	2
grok-4.20-beta1	3
gemini-3.1-pro-prev…	4
gemini-3-pro	5

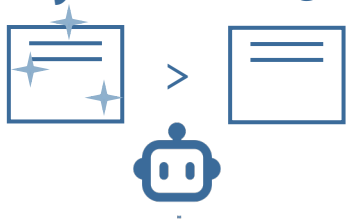
Motivation

Bias in pairwise preferences

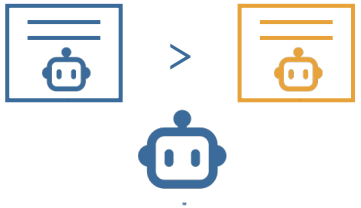
length-bias [1]

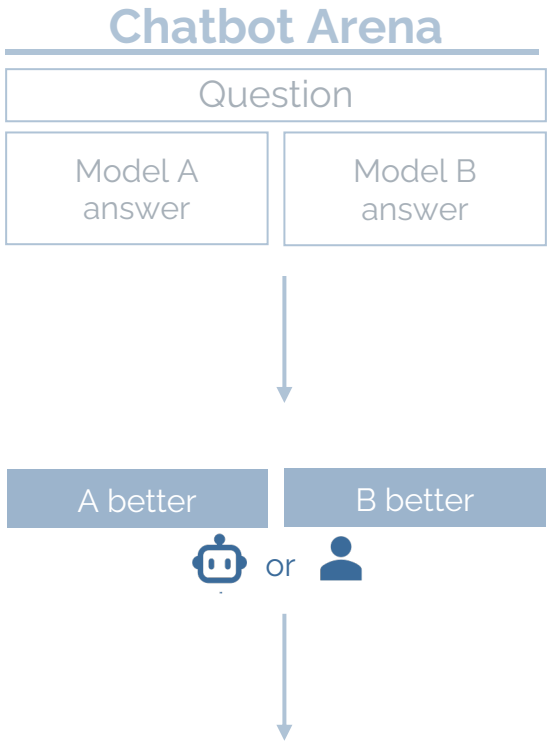


style-bias [2,3]



self-preference bias [4]





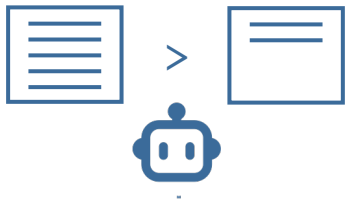
- What do pairwise preferences even measure?
  - objective differences in capabilities
  - subjective biases

[1] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto, "Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators," Mar. 10, 2025, *arXiv*: arXiv:2404.04475. doi: 10.48550/arXiv.2404.04475.
[2] T. Li *et al.*, "From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline," Oct. 14, 2024, *arXiv*: arXiv:2406.11939. doi: 10.48550/arXiv.2406.11939.
[3] G. H. Chen, S. Chen, Z. Liu, F. Jiang, and B. Wang, "Humans or LLMs as the Judge? A Study on Judgement Bias," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 8301–8327. doi: 10.18653/v1/2024.emnlp-main.474.
[4] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," Dec. 24, 2023, arXiv: arXiv:2306.05685. doi: 10.48550/arXiv.2306.05685.

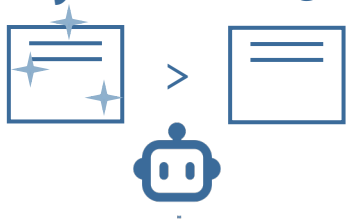
Motivation

Bias in pairwise preferences

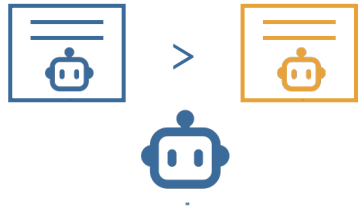
length-bias [1]

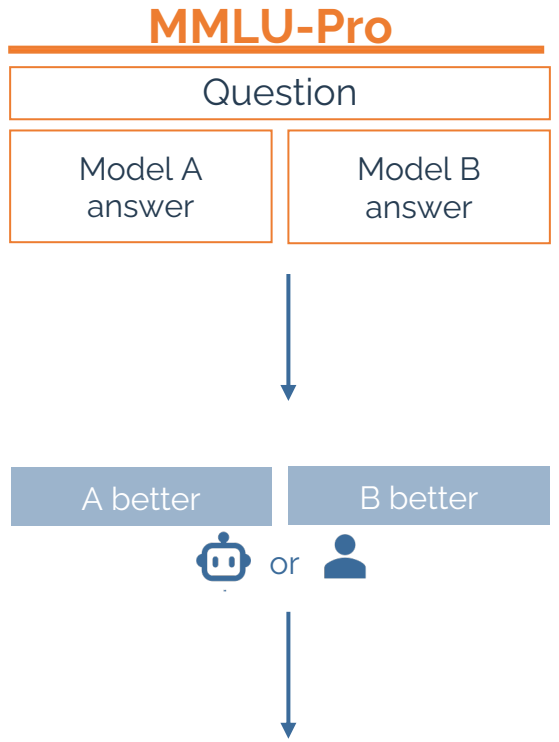


style-bias [2,3]



self-preference bias [4]





Research Question

Can pairwise preference-based rankings reflect task performance?

Method

01

Approach

Method
Pairwise preferences: LLM judge
Aggregation: Bradley-Terry

Metrics
Rank distance / correlation
Score correlation

02

Baselines

Gold standard
Accuracy-based ranking

Baseline
Direct LLM-as-a-judge

03

Setup

5 benchmarks
MMLU-Pro, GPQA Diamond, SimpleQA, GSM8K, BBH

10–27 models to rank

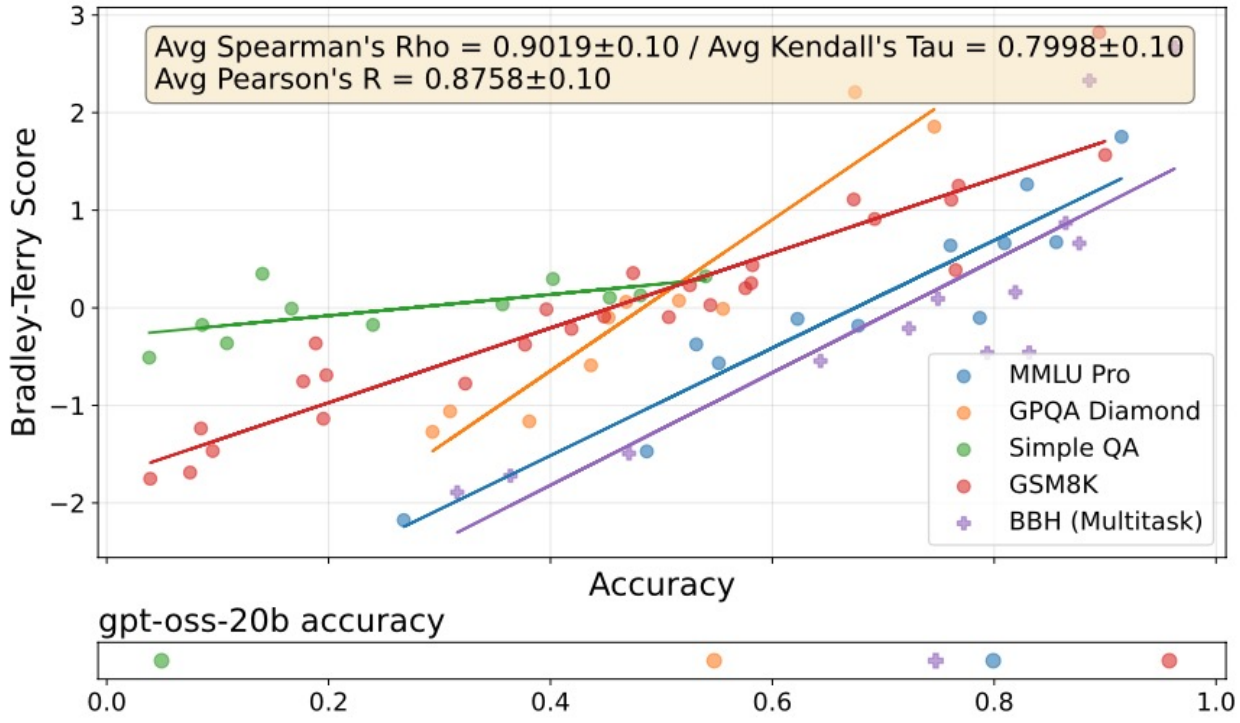
5 LLM judges
gpt-oss-20b, gpt-oss-120b, o3, phi-4, gemma-3-27b-it

Results

Agreement with accuracy-based ranking

All benchmarks
High rank correlation (Kendall's Tau >0.79)
High score correlation (Pearson's R >0.87)

Individual benchmarks
On 4 out of 5 benchmarks only 6-9% of model pairs need to be swapped to recover correct ranking.



Accuracy vs Bradley-Terry based ranking

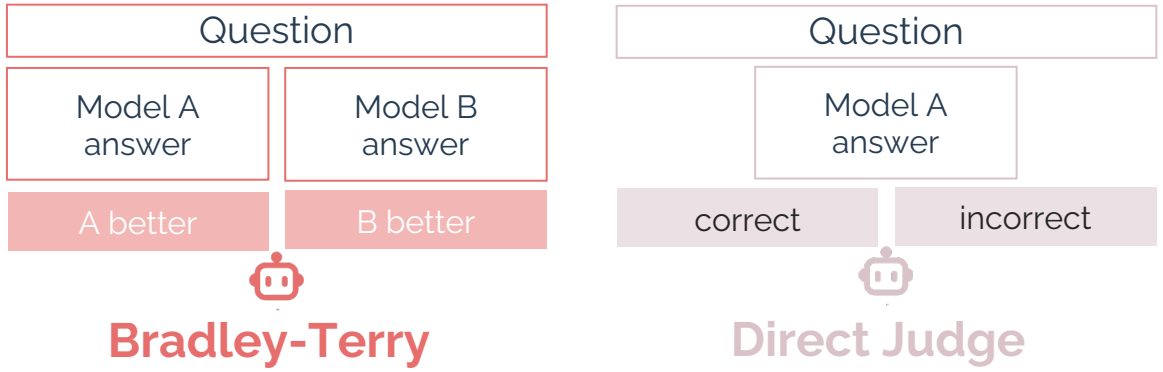
Results

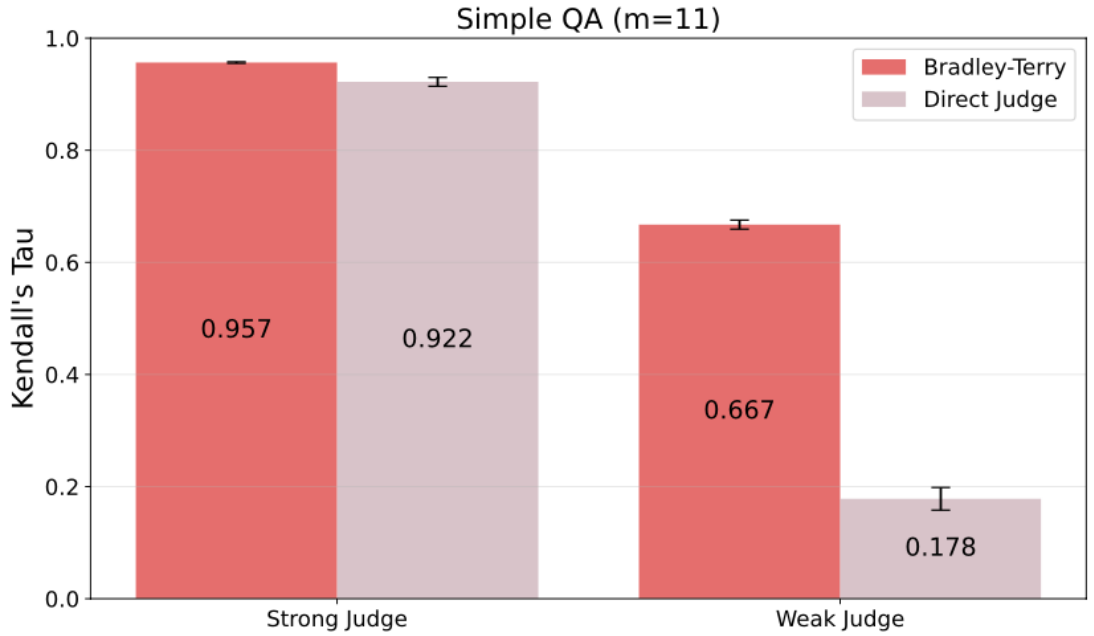
Comparison to Direct Judge baseline (SimpleQA)

Strong Judge
Good model on the task → good judge.
Overall similar results.

Weak Judge
Bradley-Terry beats the Direct Judge approach!
Bradley-Terry *halves* the rank distance ($K_D = 38.2\% \rightarrow 20\%$)

In frontier settings, it's difficult to say whether we selected a "weak" or a "strong" judge.





Rank correlation with Bradley-Terry vs Direct Judge

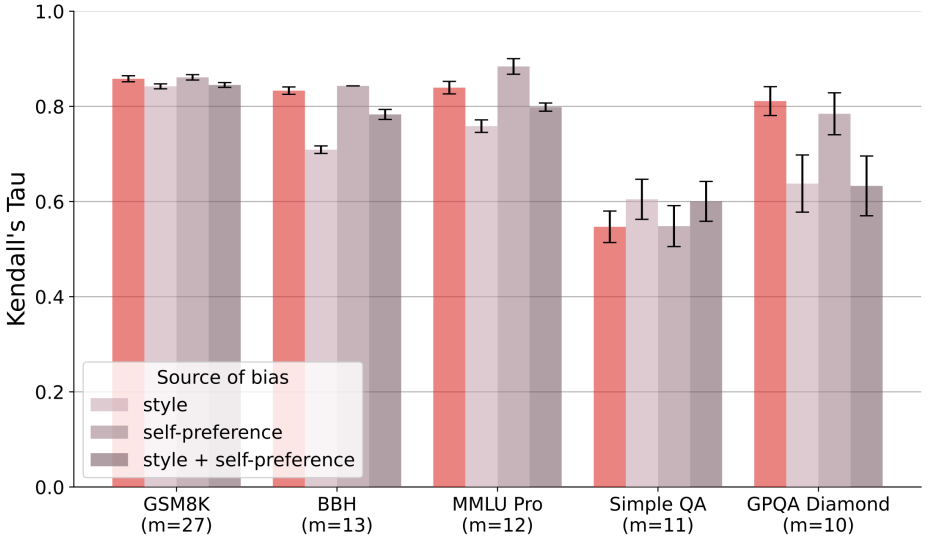
Results

The role of appearance – bias or signal?

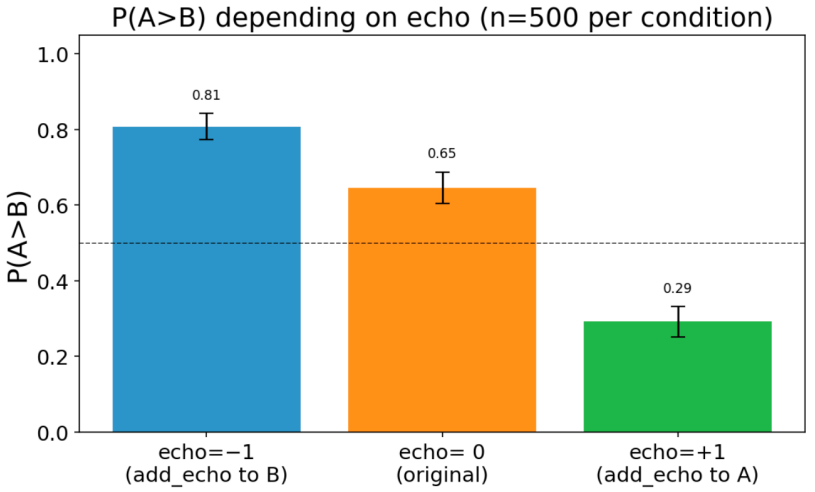
Correcting for style & self-preference
Rankings robust to style and self-preference bias.

Non-discriminative pairs *(both answers correct / incorrect)*
Affects 60% of pairs: judges rely on superficial cues.
Surprising amount of useful signal (Tau 0.8 → 0.68).

Echo *(repetitive sequence generation)*
Significant effect on judge preference: $P(A > B)$ drops from 0.81 to 0.29 ($\Delta = 0.52$).
Effect *disappears* on discriminative pairs.



Rank correlation after correcting for bias



Judge preference depending on echo

Results — Summary

01

Agreement

Pairwise preference rankings closely match accuracy-based rankings across all five benchmarks.

*Rank correlation **Tau >0.79**, score correlation **R >0.87**.*

02

Robustness

When the judge is weak, pairwise comparisons outperform the Direct Judge approach.

*Bradley-Terry **halves** the rank distance.*

03

Style & Bias

Rankings robust to style and self-preference bias. Surprising amount of signal in non-discriminative pairs.

***Echo** as a causal driver in non-discriminative pairs.*

Research Question

Can pairwise preference-based rankings reflect task performance?

→ **In this verifiable setting, yes!**